

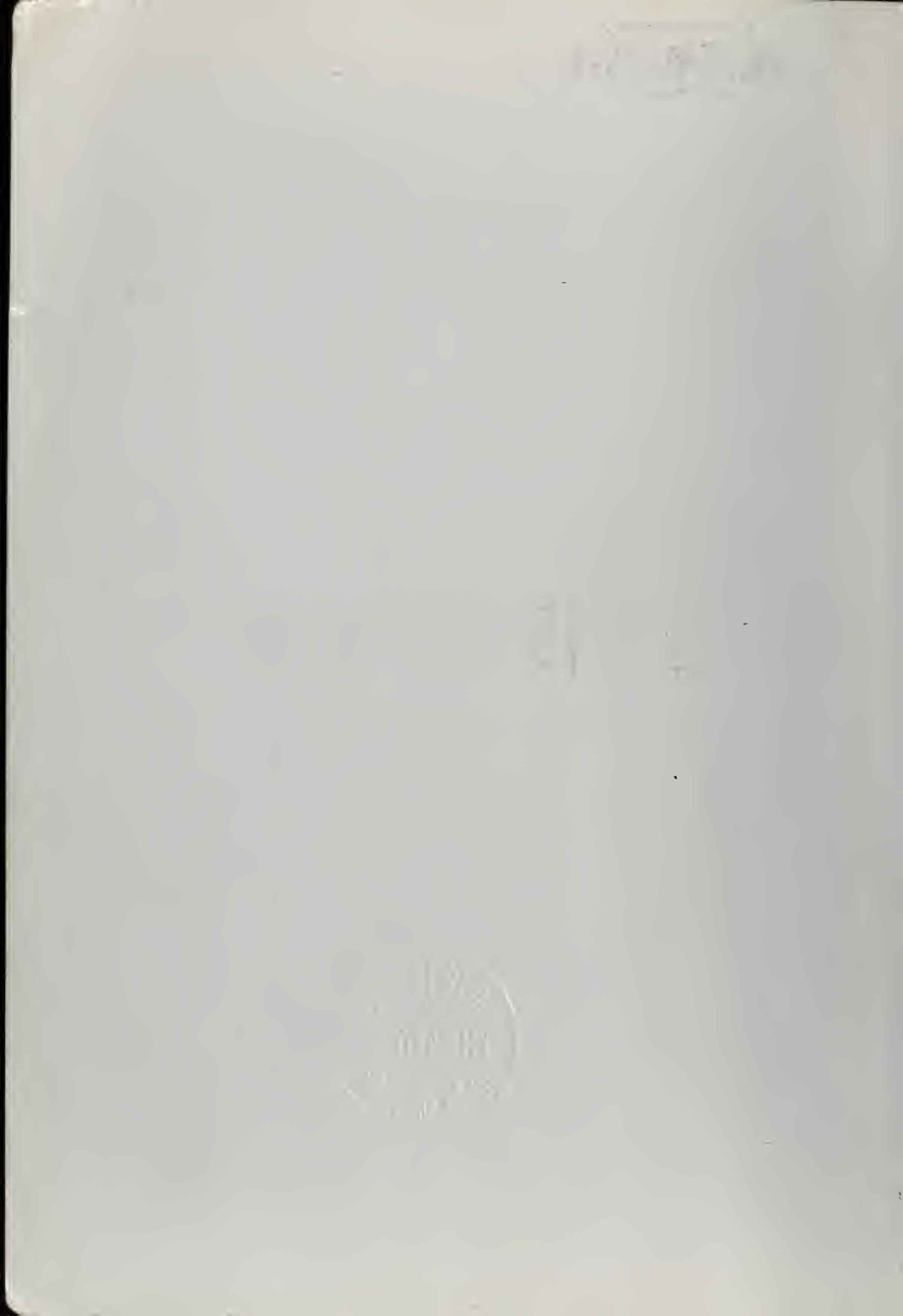
现代生物技术前沿

〔美〕D. C. 利布莱尔 著
张继仁 译
高友鹤 校

蛋白质组学导论

——生物学的新工具

科学出版社
www.sciencep.com





现代生物技术前沿

〔美〕D. C. 利布莱尔 著

张继仁 译

高友鹤 校

蛋白质组学导论

——生物学的新工具

科学出版社

北京

图字：01-2004-2025

内 容 简 介

本书介绍了分析蛋白质和肽的各种方法，重点阐述了不同质谱仪及相关数据库检索算法的基本原理和使用方法，详细描述了质谱在蛋白质组学中的应用。蛋白质组学领域最权威的科学家之一 John R. Yates 教授称本书是蛋白质组学的极好的导论和综述，可供有生物化学背景的学生和科学家使用，书写流畅，深入浅出。

本书可用作从事生命科学和医学研究的专业人员的参考书，也可用作学习生命科学和生物技术的本科生和研究生的教材。

The original English language work has been published by HUMANA PRESS, Totowa, New Jersey , U. S. A.

图书在版编目（CIP）数据

蛋白质组学导论：生物学的新工具/（美）利布莱尔（D. C. Liebler）著. 张继仁译. 一北京：科学出版社，2005
（现代生物技术前沿）
ISBN 7-03-014258-6

Ⅰ. 蛋… Ⅱ. ①利…②张… Ⅲ. 蛋白质-研究 Ⅳ. Q51

中国版本图书馆 CIP 数据核字（2004）第 098071 号

责任编辑：莫结胜 丁顺华 卢庆陶 / 责任校对：陈丽珠
责任印制：钱玉芬/ 封面设计：王 浩 陈 敬

科学出版社 出版
北京东黄城根北街16号
邮政编码:100717
http://www.sciencep.com
双青印刷厂 印刷
科学出版社发行 各地新华书店经销
*
2005年1月第 一 版 开本：B5（720×1000）
2005年1月第一次印刷 印张：8 1/4
印数：1—3 000 字数：154 000

定价：30.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

译者的话

在后基因组时代，生物学家正在对基因组的功能进行研究。蛋白质组学是生物技术的最新领域，它将对基因组功能的研究做出巨大贡献。基因只含有制造蛋白质的指令，而细胞的各种各样的生理功能是由蛋白质来完成的。蛋白质组学将组织或细胞中的蛋白质作为一个系统来研究，而不像在蛋白质化学中那样只研究蛋白质的单一组分。在不同的细胞中会有不同的蛋白质的表达，在相同细胞的不同状况中（如处在疾病状态的细胞中），也会有某些不同的蛋白质的表达。细胞只有一个基因组，但可能有许多不同的蛋白质组。蛋白质组比基因组更复杂。这种复杂性还表现在蛋白质有很多难以预测的翻译后修饰、蛋白质-蛋白质相互作用以及折叠成各种形状的三维结构。从蛋白质的氨基酸序列不一定能推测蛋白质的功能，而蛋白质组学的研究可以揭示蛋白质的表达和功能。各国科学家目前正在用蛋白质组学对人类全部蛋白质进行分类，研究人类蛋白质的相互作用，以便开发更有效的药物。

蛋白质组学面临的挑战是必须研究复杂的蛋白质体系。这要求我们分析各种各样的蛋白质，这些蛋白质大部分以修饰的形式存在并且是低丰度的。《蛋白质组学导论——生物学的新工具》一书描述了应对这个巨大挑战的工具和方法。本书介绍了分析蛋白质和肽的各种方法，重点阐述了不同质谱仪及相关数据库检索算法的基本原理和使用方法，详细描述了质谱在蛋白质组学中的应用。本书作者 D. C. 利布莱尔（Daniel C. Liebler）教授是有丰富蛋白质组学研究和教学经验的科学家，尤其在使用质谱技术及相关算法鉴定蛋白质方面有很高造诣。他特别重视蛋白质组学的应用，发表了一系列用质谱技术研究蛋白质修饰的文章。作者力图使本书成为一本可供有生物化学背景的学生和科学家使用的入门教材，书写流畅，深入浅出。蛋白质组学领域最权威的科学家之一 J. R. 耶茨（John R. Yates）教授称本书是蛋白质组学的极好的导论和综述。

本书共分三部分。第Ⅰ部分用两章描述了蛋白质组学和蛋白质组的定义，阐述了蛋白质组学诞生和发展的基础以及在新生物学中的地位，讨论了蛋白质组与基因组的关系。第Ⅱ部分介绍了蛋白质组学的工具和方法。在第 4 章和第 5 章讨论了蛋白质和肽的分离方法以及蛋白质的消化技术。蛋白质和肽的分离包括使用二维 SDS 聚丙烯酰胺凝胶电泳（2D-SDS-PAGE）、制备等电聚焦、HPLC、串联液相层析和毛细管电泳。第 6 章详细描述了 MALDI-TOF 质谱仪和 ESI 串联质谱仪的结构和工作原理以及它们的优缺点，讨论了在蛋白质组学研究中如何选用不同的质谱仪。第 7 章的主题是用肽质量指纹谱鉴定蛋白质，讨论通过测定的

肽质量与数据库理论肽质量的比较进行蛋白质鉴定的方法。第 8 章到第 10 章主要阐述如何用串联质谱分析肽序列，描述了从串联质谱谱图鉴定蛋白质的软件工具（主要介绍 Sequest），也介绍了如何使用 SALSA 算法采集串联质谱数据特征。第Ⅲ部分详细介绍了质谱和相关技术在蛋白质组学中的应用，描述了在蛋白质采集和在蛋白质表达谱的研究中，2D-SDS-PAGE 和 MALDI-TOF 质谱以及肽的多维层析和 LC-串联质谱分析的应用（第 11 章、第 12 章）；讨论了在鉴定蛋白质-蛋白质相互作用和蛋白质复合物以及鉴定蛋白质修饰中，质谱以及 Sequest 和 SALSA 算法的应用（第 13 章、第 14 章）。最后一章指出了蛋白质组学的新的发展方向，包括新质谱仪、自动化和蛋白质微阵等。

相信本书将为从事生命科学和医学研究的专业人员，以及学习生命科学和生物技术的本科生和研究生助一臂之力。

张继仁

序

自 1958 年克劳斯·比曼（Klaus Biemann）教授首先用质谱仪分析氨基酸以来，质谱技术已有了长足发展。Biemann 最初的实验所面临的棘手问题是怎样将非极性分子引入质谱仪产生离子。1958 年以后出现的几种新型电离技术和样品导入方法，促进了生物分子的分析，如化学电离、电场解吸、场致电离、等离子体解吸以及快原子轰击（FAB）等新型电离技术，鉴定肽和蛋白质的方法也得以发展。1987 年由于在生物分子中引入了基质辅助激光解吸电离（MALDI）以及电喷雾电离（ESI），质谱技术有了跃进。这两种电离技术也给肽和蛋白质分析带来极大飞跃，其中一个关键质谱技术是串联质谱。

在 20 世纪 80 年代早期唐纳德·亨特（Donald Hunt）教授开始在肽和蛋白质序列分析中发展和应用串联质谱。FAB 是一项软电离技术，可产生完整的质子化分子，使得用于肽序列分析的方法得以改进。FAB 是肽序列测定的主要突破，该技术可以使肽稳定电离，无需通过其他方法增加肽的挥发性。FAB 与串联质谱的联用，形成了快速肽序列测定方法学。处理复杂肽混合物时，大多数方法采用离线 HPLC 进行分离。人们通过这种方法对许多蛋白质进行了序列测定，并发展了许多重要的方法。然而分离方法与 FAB 的在线结合一直未能发展出可靠易行的方法。直到电喷雾电离使分离技术与质谱仪直接联用，这个问题才得到解决。分析灵敏度的增加以及样品处理的简化和自动化使肽和蛋白质分析的各个方面都得以提高。

质谱的这些进展与全球协同进行的人类基因组序列测定很好地衔接在一起。基因组序列测定工作包括人类基因组及许多模式生物的基因组，并已产生大量的序列信息。1993 年，几个研究小组发现质谱数据可用来检索数据库，以鉴定所研究的蛋白质。1994 年发展了用串联质谱数据检索序列数据库的方法，使研究者能在“书的后面看到答案”。如果“书”是已得到序列分析的生物基因组，答案基本上肯定是在书后面的部分。翻译后修饰和氨基酸序列改变等复杂问题可通过研究从基因组序列推出的蛋白质序列得以解决。

20 世纪 90 年代在生物科学中人们对质谱的兴趣和应用迅速增长，质谱在新千年将会像 SDS-PAGE 一样普遍和重要。生物学家将依赖质谱判断其实验结果。如果生物学家需要使用质谱技术来分析实验，那么他们怎样了解质谱艺术和蛋白质组学的方法呢？D. C. 利布莱尔（Daniel C. Liebler）教授的《蛋白质组学导论——生物学的新工具》一书可以指导我们了解质谱并且在蛋白质组学研究中使用质谱。这本书描述了通常使用的质谱仪和基本的电离技术，这对于确定特定研

究中如何选用质谱仪的类型是重要的。对于非专业研究人员来说，使用质谱数据检索数据库是重要的改进，这样就不再需要掌握解释质谱图的技巧。本书描述了对基础检索算法的基本理解，并阐述了其局限性，最后描述了质谱在蛋白质组学中的应用。本书为研究生和所有对迅速发展的蛋白质组学基础知识感兴趣的生物学家提供了极好的蛋白质组学导论和综述。

J. R. 耶茨（John R. Yates）
Scripps Research Institute
La Jolla，CA

前　言

本书是蛋白质组学这个新领域的导论，侧重描述怎样分析研究蛋白质和蛋白质组。尽管人们对蛋白质组学的兴趣日益浓厚，但是对蛋白质组学的工具和技术了解还很少。本书注重向生物学家介绍新工具和新方法，对生物学学生和有经验的生物学家都适用。任何学过研究生生物化学课程的人都可以很容易地理解什么是蛋白质组学以及如何研究蛋白质组。有经验的生物学家会发现本书大部分内容是熟悉的，但是这些内容被重新整合并围绕蛋白质组的研究展开阐述。

基因组序列测定、分析仪器、计算能力和易于使用的软件工具等方面的重大进展已不可逆地改变了生物学的发展方向。过去我们一直研究生物系统的单个组分，而现在可以综合地并且在精确的分子细节上研究生物系统本身。我们面对的任务是有效地利用新技术和处理大量的数据，更重要的是我们需要调整思想去理解与单一组分相对的复杂体系。

《蛋白质组学导论——生物学的新工具》这本书最早是用质谱进行肽序列分析的短期课程讲义，这门课程是由 Donald F. Hunt 博士 1998 年在北卡罗来纳州 Durham 的生物医学资源设施协会会议上讲授的。那时我的同事 Tom McClure 博士和我在亚利桑那大学毒理中心和亚利桑那癌症中心建立了一个新的蛋白质组学研究机构。Tom 参加了 Hunt 的课程。他回来后，将授课内容传授给我们几个。我们于 1998 年 8 月在亚利桑那大学开设了一个为期四天的关于质谱和蛋白质组学的训练班，共有 50 个学员接受了培训，学员包括研究生、实验室工作人员和教授。对这个训练班的热烈反响反映了以某些易接受方式将蛋白质组学的新技术和在研究中的潜在应用介绍给科学家的需要。这个训练班促进了本书的诞生。

本书是写给初学者的。我的目的是让他们熟悉蛋白质组学的重要工具和应用，所以对某些仪器和应用的描述并不是非常严谨的。本书不是实验室手册或最新技术汇编。有几本很好的书更详细地描述了蛋白质分析技术、质谱仪器和技术，以及这些技术的应用。在这个领域中研究方法的发展和应用是非常迅速的，没有哪本书是真正时新的。在我将蛋白质组学介绍给同事后令我兴奋的是同事们创造性地运用这些新技术，这将促进蛋白质组学的发展。

本书分成三部分。第Ⅰ部分介绍蛋白质组学主题，描述它在新生物学中的位置，分析蛋白质组的性质。第Ⅱ部分介绍蛋白质组学研究的工具，解释它们怎样工作。第Ⅲ部分解释这些工具怎样用来解决不同类型的生物学问题。

感谢 Jeanne Burr、Laura Tiscareno、Julie Jones、Dan Mason、Beau Hansen、Hamid Badghisi、Linda Manza、Richard Vaillancourt、Tom McClure、Arpad Somogyi 和 George Tsaprailis，他们提出了很好的建议，阅读书中各章的草稿并给出评语，提供某些图解的样品数据。感谢 Elizabeth Hedger 杰出的秘书工作。最后，感谢我的妻子 Karen 和儿子 Andrew 对我写作的支持。

D. C. 利布莱尔（Daniel C. Liebler），PhD

目　录



I　蛋白质组学和蛋白质组



1 蛋白质组学和新生物学

1.1 新生物学

蛋白质组学是研究与基因组相对应的蛋白质组的学科。术语“蛋白质组学”（proteomics）和“蛋白质组”（proteome）是 Marc Wilkins 及其同事在 20 世纪 90 年代早期提出的，对应于描述生物中全部基因的术语“基因组学”（genomics）和“基因组”（genome）。这些“-omics”术语代表了对怎样思考生物学和生物体系工作方式的重新定义（图 1.1）。直到 20 世纪 90 年代中期，生物化学家、分子生物学家和细胞生物学家还在研究单独的基因和蛋白质或与生物化学途径相关的少量组分。那时可用的技术有 Northern 印迹法（用于基因表达分析）和 Western 印迹法（用于蛋白质分析），利用这些技术研究和分析多基因或多蛋白质是非常困难的。

三项进展形成了新生物学的基础，改变了生物学研究前景。第一项是 20 世纪 90 年代基因、表达序列标签（EST）和蛋白质序列数据库的发展。作为许多生物基因的部分信息库，这些资源很有价值。20 世纪 90 年代后期的基因组序列测定工作，阐明了细菌、酵母、线虫和果蝇的完整基因组序列，最近阐明了人类基因组完整序列。植物和其他广泛研究的动物基因组的序列最近也已完成或接近完成。这些基因组数据库是我们最终从中获取对生物体系理解的信息库。

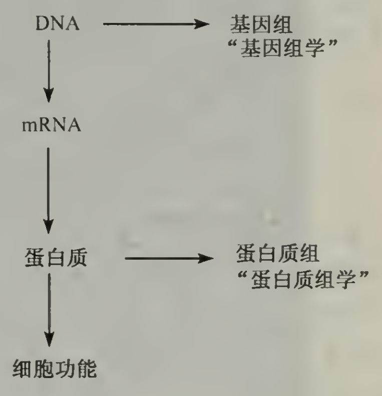


图 1.1 基因组学和蛋白质组学的生物化学关系

第二项重要进展是引入易于操作的、基于浏览器的生物信息学工具。利用这些工具从上述数据库中获取信息。现在可以在几秒钟内从完整的基因组内检索特定的核酸或蛋白质序列。这样的数据库检索工具与其他工具和数据库结合利用，根据已存在的特定功能结构域和基序可以预测蛋白质产物的功能。这样一批基于互联网的免费工具使生物学家可以通过台式电脑检测基因和基因产物的结构与功能，探索大量感兴趣的生物化学问题。

第三项重要进展是寡核苷酸微阵。微阵含有在玻片上或芯片上的一系列基因专一性寡核苷酸或 cDNA 序列。将感兴趣样品的 DNA 混合物荧光标记后与微阵进行杂交，可以一次探测几千个基因的表达。一个微阵可以代替几千个 Northern 印

迹法分析，可以在做一次 Northern 印迹的时间内完成。通过使用双色荧光探针标记，两个不同样品的基因表达谱可直接在一个玻片或芯片上进行比较。

图 1.2 是一块含有酿酒酵母基因组中 6 000 个基因单一序列的玻片。这样的一个微阵可以测定酵母基因组所有基因的表达。显然，这使我们面临新生物学的巨大挑战。我们可以观看整个系统，但这几千个数据点所包含的信息超出了我们直观解释的能力。新的组合算法、自组作图和类似的工具等最新的方法有助于生物学家理解这些数据。

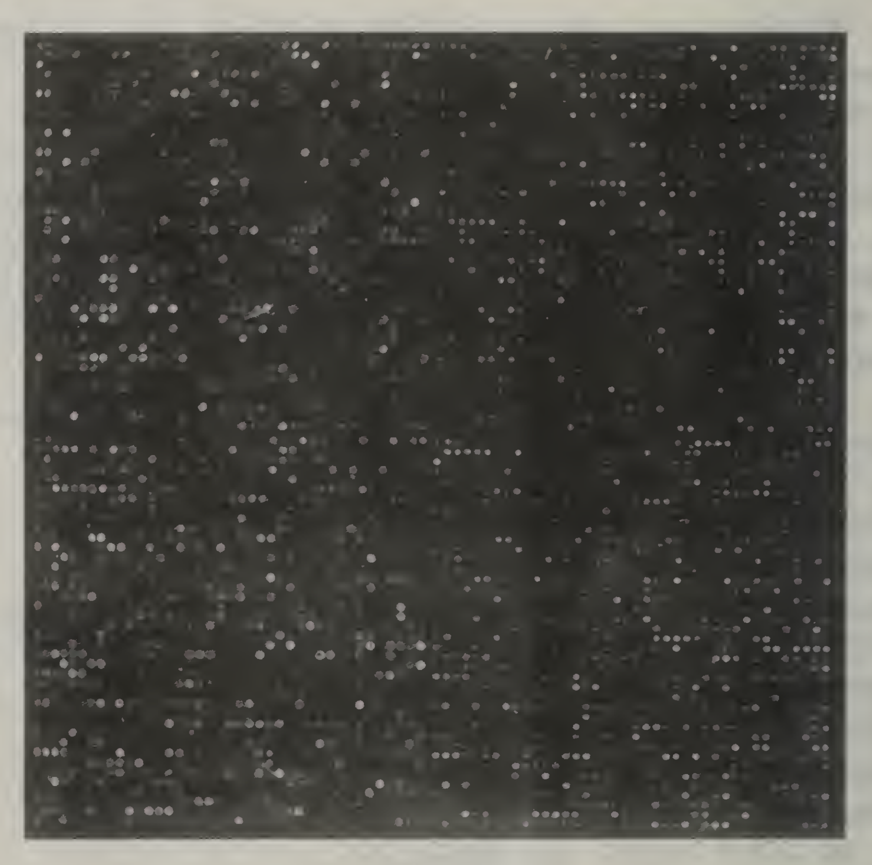


图 1.2　酵母基因组芯片

该酵母 cDNA 微阵由斯坦福大学 Patrick Brown 博士的实验室制作（http://cmgm.stanford.edu/pbrown/）。

微阵带来的最重要的变化是使生物学家进行“宏观”思考。细胞有成千上万的以不同组合方式进行的基因表达。细胞的生死是由这些基因的表达和其蛋白质产物的活性决定的。无论是跨膜受体、转录因子、蛋白激酶或是伴侣分子，每种蛋白质所表达的功能只有在同一细胞内其他蛋白质的功能和活性同时表达时才有意义。因此生物学家正在努力做宏观思考，去理解系统而不只是理解组分，寻求复杂性的意义。

1.2 蛋白质组学？它不过是我们过去称之为蛋白质化学的学科！

对新思想、术语和方法人们通常说它们其实不是新发展来的，因此解释蛋白质组学与蛋白质化学的不同是很重要的。表 1.1 对各自的主要特征作了小结。蛋白质化学包括研究蛋白质的结构和功能，通常涉及物理生物化学或机械酶学。研究工作通常包括完整序列测定、结构测定以及进行结构控制功能的模型研究。物理生物化学家和酶学家在同一时间内只研究一个蛋白质或多亚基蛋白质复合物。

表 1.1　蛋白质化学和蛋白质组学的不同

蛋白质化学	蛋白质组学
单一蛋白质	复杂混合物
全序列分析	部分序列分析
强调结构与功能	强调通过数据库匹配进行蛋白质鉴定
结构生物学	系统生物学

蛋白质组学研究多蛋白质系统，重点研究作为一个大系统或部分网络的组成的多个不同蛋白质的相互作用。蛋白质组学需进行复杂混合物的分析，不是通过完整序列测定进行鉴定，而是在数据库匹配工具帮助下进行部分序列测定。蛋白质组学的内容是系统生物学，而不是结构生物学。换句话说，蛋白质组学的要点是鉴定系统的行为而不是任何单一组分的行为。

1.3 我们能测定基因表达，为什么还要有蛋白质组学？

基因微阵提供了细胞中大量或全部基因表达的快速检测。然而从 mRNA 水平并不一定能预测细胞中相应蛋白质的水平。各种 mRNA 不同的稳定性和不同的翻译效率能够影响新蛋白质的产生。蛋白质形成后，在稳定性和转换速度上有很大不同。许多参与信号传导、转录因子调节和细胞周期控制的蛋白质迅速转换，这是其活性调节的一种方式。mRNA 水平没有告诉我们相应蛋白质的调节状态，蛋白质的活性和功能常有一些内源翻译后的改变，也会因环境因素而改变。

1.4 蛋白质组学：对分析的挑战

如何同时测定一个生物中大量或全部基因的表达似乎已通过引入 cDNA 或寡核苷酸微阵得以解决。用 DNA 微阵和相关方法分析基因表达依赖于两个重要工具：聚合酶链反应（PCR）和寡核苷酸与互补序列的杂交。但是没有类似的工具用于蛋白质分析。首先，没有蛋白质 PCR 等价物。目前不可能有多肽分子以类似于核苷酸通过 PCR 复制的方式复制。少量的寡核苷酸可以通过 PCR 进行扩增，而少量的蛋白质必须在没有任何扩增的情况下进行测定和分析。

第二，蛋白质不能专一性与互补氨基酸序列杂交。Watson-Crick 碱基配对

允许寡核苷酸与互补序列杂交。一个特定的互补寡核苷酸序列可以作为高度专一性探针，一个特定的 mRNA 或其他核酸片段可以与之结合。这种专一性允许在微阵上有一个特定的点以便识别单一序列。尽管抗体和寡核苷酸结合子（aptamer，也称适体）可以识别特定的肽或蛋白质，但是这种识别不能简单地根据序列来预测，而寡核苷酸的杂交则可以根据序列来预测。

另一个蛋白质组学的特有问题是细胞中每一个蛋白质产物并不一定只有一种分子实体。这是由于蛋白质有翻译后修饰。修饰的内容和变化随不同的蛋白质、细胞的调节机制和环境因子而变化，许多蛋白质以多种形式存在。对任何特定基因的多种蛋白质产物进行检测和区分的必要性使蛋白质组学在分析方面更具挑战性。

蛋白质组的分析需要一套不同于基因表达分析的工具，能够对修饰和非修饰的蛋白质进行检测和定量的分析。我们怎样应对这项任务是这本书的主题。

1.5 蛋白质组学的工具

尽管上面描述了分析蛋白质组学的不利条件，但是鉴定蛋白质组及其组分实际上可以完成。这是由于以下 4 种重要工具的发展和结合使用给研究人员提供了灵敏性和专一性较高的识别和鉴定蛋白质的方法。

第一种工具是数据库。蛋白质、EST 和基因组序列数据库共同提供了生物表达全部蛋白质的完整数据库目录。例如，根据对果蝇的所有编码序列的分析，我们知道有 110 个果蝇基因编码具有 EGF 类结构域的蛋白质，87 个基因编码具有酪氨酸激酶催化结构域的蛋白质。在进行果蝇蛋白质组学研究时，我们检索大量已知的可能蛋白质结构域。当用限定的序列信息，甚至原始质谱数据（见下文）进行检索时，我们可以根据质谱数据与数据库的匹配情况鉴定蛋白质组分。

第二种工具是质谱（MS）。质谱仪的使用在过去十年中有了极大的革新，在发展为分析生物分子，特别是分析蛋白质和肽的高灵敏度和高可靠性上达到顶点。MS 仪器的使用可提供三类分析，这三类分析在蛋白质组学分析中都非常有用。首先，MS 可以进行 100 kDa 或更大完整蛋白质的精确质量测定。估计蛋白质质量的最好方法是 MS 分析，而不是测定蛋白质在十二烷基硫酸钠-聚丙烯酰胺凝胶电泳（SDS-PAGE）的迁移。高精度蛋白质质量测定的应用有限，因为它们往往不够灵敏。净质量对精确鉴定蛋白质往往是不充分的。其次，MS 也能对蛋白质水解消化产生的肽进行精确的质量测定。相对于完整蛋白质质量测定，肽质量测定可以有高灵敏度和高质量准确度。可以直接用肽质量测定数据在数据库中进行检索，这样常常可以确切鉴定靶蛋白质。最后，MS 可以对蛋白质水解消化得到的肽序列进行分析。目前认为 MS 是肽序列分析中的最新技术。MS 序列数据为蛋白质鉴定提供了最有力和最精确的方法。

蛋白质组学的第三个必要工具是对数据库中特定蛋白质序列与 MS 数据进行比对的各种软件。前面提到从 MS 数据可以测定序列，但是这种从头开始分析成

百上千的谱图时是一项费时费力的任务。蛋白质组学软件将未分析的MS数据，在特定算法的帮助下与蛋白质、EST和基因组序列数据库的序列相比对，自动检测大量用于蛋白质序列匹配的MS数据。然后研究人员检查自动检测的结果，估计数据的质量，所用的时间比手工解释每一张谱图要少得多。

蛋白质组学的第四种必需工具是蛋白质的分析分离技术。在蛋白质组学中蛋白质分离有两个目的。第一，通过将蛋白质混合物分离成单一蛋白质或蛋白质小组以简化复杂蛋白质混合物。第二，蛋白质的分离分析可以比较两个样品蛋白质的不同表观，研究者可以标记用于分析的特定蛋白质。2D-SDS-PAGE是最广泛用于蛋白质组学的技术。2D凝胶电泳也许是在复杂样品中分离蛋白质的最好单项技术。其他的蛋白质分离技术，包括1D-SDS-PAGE、高效液相层析（HPLC）、毛细管电泳（CE）、等电聚焦（IEF）和亲和层析，也都是分析蛋白质组学的有用工具。最有力的技术是将不同的蛋白质和肽分离技术结合为多维技术。例如，离子交换液相层析（LC）与反相（RP）-HPLC的串联是分离复杂肽混合物的有力工具。

这四种工具的结合形成了蛋白质组学当前的技术，每一种工具在技术上都发展迅速。在本书的后面几章我们将讨论每一种分析工具。

1.6　蛋白质组学的应用

蛋白质组学技术确实很新颖，但是鉴定蛋白质组究竟是为了什么呢？根据目前的实践，蛋白质组学包括4项主要应用，它们是：①采集；②蛋白质表达谱；③蛋白质网络谱；④蛋白质修饰谱。下面对上述每一项进行简短定义，在本书其后各章将详细讨论。

采集是鉴定样品中所有（或尽可能多）的蛋白质。采集主要是直接对蛋白质组进行分类，而不是通过基因表达（如通过微阵）数据来推断蛋白质组的组成。蛋白质组学中采集需要耗费大量的劳动，以使蛋白质得到最大程度的分离，然后使用MS和相关的数据库以及软件工具进行鉴定。有几种采集方法，每一种都有其优点。这些方法的联用可直接分析证实那些只能从基因表达数据推断的数据。

蛋白质表达谱是鉴定生物或细胞特定状态（如分化、发育状态或疾病状态）下蛋白质的表达或药物、化学或物理刺激下蛋白质的表达。表达谱其实是特殊的采集形式，分析中比较一个特定系统的两种不同状态。例如，比较正常细胞和病理细胞中哪些蛋白质有不同的表达。这种信息对于检测药物治疗的潜在靶子极为有用。

蛋白质网络谱是在生物系统中测定蛋白质之间相互作用的蛋白质组学方法。大多数蛋白质在执行功能时与其他蛋白质密切相关。这些相互作用决定蛋白质功能网络（如信号传导级联过程和复杂的生物合成或降解途径）的功能。大多数蛋白质-蛋白质相互作用是通过体外纯化的蛋白质和用酵母双杂交系统获得。通过亲和俘获配对技术与分析蛋白质组学方法相结合，蛋白质组学可以鉴定更复杂的

蛋白质网络。蛋白质组学方法已用来鉴定多蛋白质复合物的组分。在细胞中多蛋白质复合物与点到点的信号传导途径有关。蛋白质网络谱可以测定途径中所有参与者的状态。蛋白质网络谱是蛋白质组学最具远大前景的应用之一。

蛋白质修饰谱是鉴定蛋白质怎样和在何处被修饰的。许多蛋白质翻译后的修饰控制着蛋白质的靶向、结构、功能和转换。此外，许多环境化学因素、药物和内源化学因素可产生修饰蛋白质的活性亲电体。研究人员已开发了各种分析工具用以鉴定修饰蛋白质和修饰的性质。修饰蛋白质可用抗体测定（如用特定磷酸化氨基酸残基的抗体），但是一个特定修饰的精确序列位点往往是未知的。蛋白质组学方法是研究翻译后修饰的性质和序列专一性的最好方法。这项方法的扩展允许在一个网络中同时鉴定调节蛋白质的修饰状态，这是蛋白质组学技术的重要扩充。这些方法用新方式回答蛋白质组的化学修饰怎样影响生物系统的问题。

推荐读物

Brown, P. O. and Botstein, D. (1999) Exploring the new world of the genome with DNA microarrays. *Nat. Genet.* **21**, 33－37.

DeRisi, J. L., Iyer, V. R., and Brown, P. O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278**, 680－686.

Eisen, M. B., Spellman, P. T., Brown, P. O., and Botstein, D. (1998) Cluster analysis and display of genome-wide expression patterns. *Pros. Natl. Acad. Sci. USA* **95**, 14, 863－14, 868.

Fields, S. (2001) Proteomics in genomeland. *Science* **291**, 1221－1224.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860－921.

Lashkari, D. A., DeRisi, J. L., McCusker, J. H., Namath, A. F., Gentile, C., Hwang, S. Y., et al. (1997) Yeast microarrays for genome wide parallel genetic and gene expression analysis. *Proc. Natl. Acad. Sci. USA* **94**, 13, 057－13, 062.

Pandey, A. and Mann, M. (2000) Proteomics to study genes and genomes. *Nature* **405**, 837－846.

Venter, J. C., Adams, M. D., Myers, E. M., Li, P. W., Mural, R. J., et al. (2001) The sequence of the human genome. *Science* **291**, 1304－1351.

Wilkins, M. R., Sanchez, J. C., Gooley, A. A., Appel, R. D., Humphery-Smith, I., Hochstrasser, D. F., and Williams, K. L. (1996) Progress with proteome projects: why all proteins expressed by a genome should be identified and how to do it. *Biotechnol. Genet. Eng. Rev.* **13**, 19－50.

2 蛋白质组

2.1 蛋白质组与基因组

每一个人类细胞中含有制造一个完整的人所需的全部信息。然而，并不是全部基因在所有细胞中都表达。编码细胞实现基本功能（如葡萄糖代谢、DNA 合成）必需酶的基因在所有细胞中表达，而具有高度专一功能的基因只在特定类型的细胞中表达（如视紫红质在视网膜色素上皮细胞中表达）。一个细胞中表达两类基因：①必需功能蛋白质的基因；②行使细胞专一性功能蛋白质的基因。因此，一种生物有一个基因组，但有许多蛋白质组。

任何细胞的蛋白质组是所有可能基因产物的某种子集，但这并不意味着蛋白质组比基因组简单。事实正相反，任何蛋白质，即使只是同一个基因的产物，也可能存在多种形式。在一个特定的细胞内或在不同的细胞之间，蛋白质的存在形式可能不同，大多数蛋白质都以几种不同的修饰形式存在。这些修饰影响着蛋白质的结构、定位、功能和转换。

在这一章从 5 个方面讨论蛋白质组。首先，扼要讨论蛋白质的“生命周期”，从蛋白质作为翻译产物在核糖体上出现，到翻译后的多种修饰，再到最终降解。第二，讨论可根据蛋白质的序列基序、结构域的结构和生化功能分成具有不同标准组件结构的蛋白质。第三，讨论功能蛋白质家族在基因组中的分布。第四，通过基因组序列讨论在生物系统中有多种功能和冗余功能的蛋白质组。最后，讨论在特定时间内影响一个蛋白质在细胞中存在数量的各种因素，并讨论这些因素给蛋白质组学分析方法带来的困难。

2.2 蛋白质的生命周期

蛋白质的合成是通过在核糖体上将 mRNA 翻译成多肽来进行的。大多数情况下，最初的多肽翻译产物要进行某种类型的修饰后才具有功能。这些修饰统称为“翻译后修饰”，包括各种可逆和不可逆化学反应。已报告有接近 200 种不同类型的翻译后修饰。图 2.1 表明一个原型蛋白质的生命周期，并总结了一些修饰作用。

蛋白质是 mRNA 序列通过核糖体翻译产生的。半胱氨酸的巯基形成二硫键的折叠和氧化，使多肽随机卷曲具有二级结构。若干“永久”修饰，如谷氨酸的羧化和去除N端甲硫氨酸，在多肽生成的早期发生。在高尔基体的进一步加工

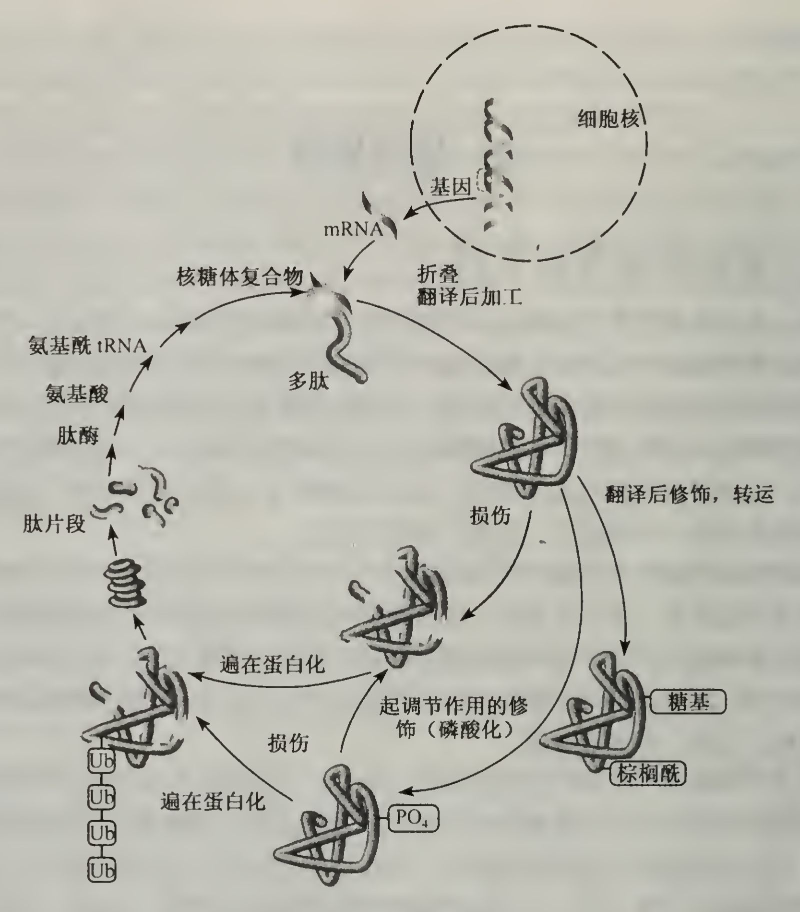


图 2.1　蛋白质的生命周期

产生糖基化。蛋白质到特定的亚细胞或细胞外区域的专一性运送往往需要有前导肽或信号肽序列，这些前导肽或信号肽在完成定位后被蛋白水解酶切割除去。某些蛋白质上还进行辅基加成作用。一个特定蛋白质可与其他蛋白质结合形成多亚基复合物。半胱氨酸残基的棕榈酰化或异戊二烯化辅助蛋白质进入膜内或嵌在膜上。这些不同程度的"永久"性修饰和运送使功能蛋白质进入细胞的特定位置。

蛋白质在细胞内的特定部位执行其功能。翻译后修饰调控许多蛋白质的活性，最重要的和研究较深入的翻译后修饰是丝氨酸、苏氨酸或酪氨酸残基的磷酸化。磷酸化可以使酶活化或失活、改变蛋白质-蛋白质相互作用和连接、改变蛋白质结构、引起蛋白质降解。蛋白质的磷酸化以各种形式调节蛋白质的功能，是信号传导级联反应、细胞周期调控和其他重要细胞功能的快速调控中的关键开关。

蛋白质也易受损伤。在生物系统中普遍存在的自由基和其他氧化剂导致蛋白质的氧化损伤。半胱氨酸巯基对氧化特别敏感。甲硫氨酸、色氨酸、组氨酸和酪氨酸残基容易发生氧化。氨基酸也易受脂类和糖类氧化产物的攻击，这包括活化

的 α，β-不饱和羰基化合物。除了这些引起蛋白质修饰的内源因素外，辐射、化学和药物等环境因素也能引起蛋白质的共价修饰或氧化修饰。许多修饰作用可引起蛋白质失活。各种修饰因素都可改变某些蛋白质的结构。

蛋白质修饰对于引发蛋白质降解的过程很重要。某些蛋白质磷酸化后迅速与泛素连接，并被 26S 蛋白酶体复合物降解。细胞中其他因素也可导致蛋白质的泛素化，包括氧化损伤和蛋白质的其他修饰。蛋白质也可以通过溶酶体酶降解。

图 2.1 表明了蛋白质组的一个要点：在细胞中任何时间任何蛋白质都可能以多种形式存在，从而使蛋白质组变得异乎寻常地复杂。另一方面，蛋白质组的状态反映了细胞的所有功能状态。

2.3 具有标准组件结构的蛋白质

另一种研究蛋白质的方式是把它们想像成具有标准组件或标准组件拼接的结构。某些氨基酸序列倾向于形成如 α 螺旋或 β 折叠的二级结构，或随机卷曲结构。特定的氨基酸序列和从这些序列产生的二级结构具有特定的性质和功能。可以认为氨基酸序列片段是功能的建筑部件或组件。大自然用这些组件建造了工具箱，通过这个工具箱，可以建造多种具有相关功能的蛋白质。

具有特定性质和功能的蛋白质标准组件单位称为“基序”或“结构域”。不同蛋白质中存在的结构域，其可识别序列有类似性质或功能。一般在使用时，这些术语往往可替换使用。在某些情况下，基序或结构域的氨基酸序列高度保守，不随蛋白质的不同而变化。还有一些情况下，一个序列中某些关键氨基酸以重复形式存在，而另一些氨基酸可有各种变化。

即使某些短序列也能赋予蛋白质某些专一性修饰。例如，*N*-糖基化的蛋白质序列中多含有三肽序列“Asn-Xaa-Ser/Thr”。在这个序列中，靶子天冬酰胺之后可以是任何氨基酸，接下来的氨基酸是丝氨酸或是苏氨酸。如果 Xaa 是脯氨酸，则不能进行糖基化修饰。尽管这个三肽序列并不能保证一定进行 *N*-糖基化修饰，但是它提供了一种基序，这种基序提供可能的生物化学作用。

较长氨基酸序列常形成结构域，赋予蛋白质特定的性质或功能。某些结构域结构只是赋予一个肽段重要物理性质，如跨膜结构域，一般形成跨脂双层膜的 α 螺旋。另一些结构域能够为重要酶底物和辅基提供氢键或其他接触。例如，真核丝氨酸/苏氨酸蛋白激酶有一个富甘氨酸的核心结构域。这个富甘氨酸区域位于与 ATP 结合的赖氨酸残基和作为催化中心的保守的天冬氨酸残基周围。在许多情况下，结构域由二级结构单位的结合组成，如螺旋-环-螺旋结构域。

基序和结构域对蛋白质组的意义是它们代表从蛋白质序列到蛋白质功能的翻译。当已知性质和功能的结构域和基序出现在未知功能的蛋白质中时，可以推测其细胞功能。简言之，分析蛋白质组学可以确定序列，序列可以确定功能。

2.4 功能蛋白质家族

另一种研究蛋白质组的方式是将它分成具有类似功能的蛋白质家族。例如，某些蛋白质具有结构作用，另一些蛋白质参与信号传导途径，还有一些蛋白质调控如核酸合成或糖代谢等必需代谢途径。根据结构域及其相关功能作用的分类，Venter 及同事推测出人类基因组所编码蛋白质的功能分布（图 2.2）。

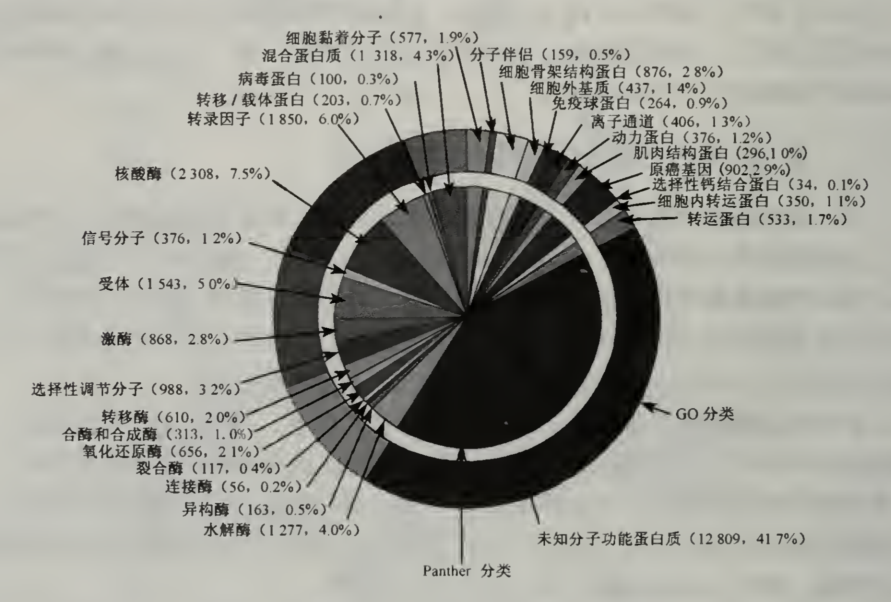


图 2.2　推测的人类基因组蛋白质产物的功能

［Venter et al.（2001）Science，291：1304～1351］

与中间代谢和核酸代谢有关的酶占蛋白质组的 15%。与结构和蛋白质合成与转换相关的蛋白质（细胞骨架蛋白、核糖体蛋白、分子伴侣和蛋白质降解相关因子）总共占 15%～20%。信号传导蛋白和 DNA 结合蛋白占 20%～25%。尽管这些数字给由蛋白质功能来解析基因组提供了有价值的参考数据，但我们并不能确定在细胞中某一特定时间某一特定蛋白质或某类特定蛋白质的表达量。接近 40%的基因组编码的蛋白质功能尚属未知。确定这些基因产物的功能是人类功能基因组学面临的最基本的挑战。

2.5 从基因组推算蛋白质组

对鉴定生物基因组的研究人员来说一个最有趣的问题是“有多少基因?”。这个问题的答案可以使我们得到共有多少蛋白质存在于蛋白质组中的概念。几种生物的全部基因组序列已经测序完成，这些数据允许分析家预测所有基因的产物。根据推测的每一个基因产物的氨基酸序列，按照蛋白序列中所含的结构域和序列

基序已对它们进行分类。例如，酿酒酵母有 119 个基因编码具有真核蛋白激酶结构域的蛋白质，另外有 47 个基因编码具有 C2H2 类型锌指结构域的蛋白质。结构域序列特征与基因组序列的比较表明生物基因组为各种类型蛋白质编码。

最近对酿酒酵母、线虫和果蝇的分析揭示出这些生物的基因组大小与预测的蛋白质组容量之间有着非常有趣的关系。Gerald Rubin 和同事根据已存在的特定的结构域，对通过流感嗜血杆菌、酿酒酵母、线虫和果蝇基因组预测的蛋白质产物进行了分类（图 2.3）。比较所有预测的蛋白质产物，结果表明基因组中存在序列与其他物种蛋白质序列稍有不同的蛋白质。通过对这些冗余蛋白质产物（称为平行进化同源物）的校正可以计算每一种生物的“核心蛋白质组”。这种核心蛋白质组代表生物的不同蛋白质家族的汇集。

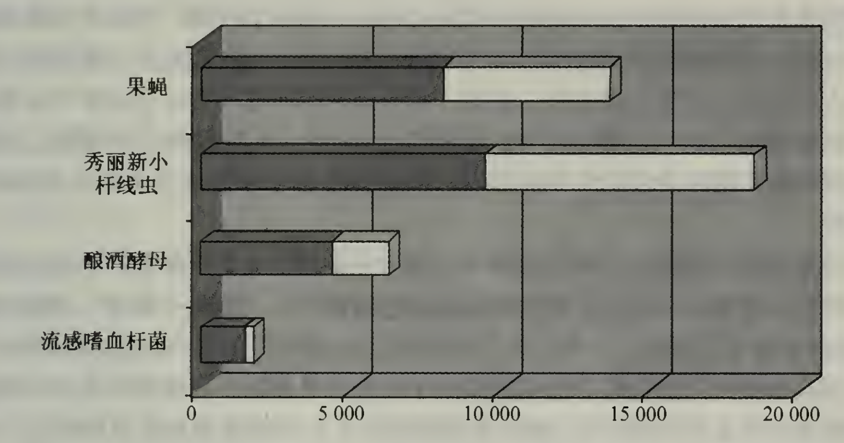


图 2.3　根据流感嗜血杆菌（1 709 个基因）、酿酒酵母（6 241 个基因）、秀丽新小杆线虫（18 424 个基因）和果蝇（13 601 个基因）基因预测的蛋白质产物

实心柱表明为单一蛋白质编码的基因，空心柱表示为平行进化同源物编码的基因。

对这些生物的核心蛋白质组的研究揭示出两个令人感兴趣的方面。首先，一种生物的复杂性和基因组中基因数量之间的关系是复杂的。当然，酵母要比细菌的基因数目多，比蠕虫和蝇类少。然而，蝇类（果蝇）是比蠕虫（线虫）要复杂得多的生物，它却有较少的基因（蝇类 13 601 个，蠕虫 18 424 个），有较少的核心蛋白质组（蝇类有 8 065 个特有蛋白质，蠕虫则有 9 543 个）。这说明生物复杂性并不是因为数量较大的基因数目，而更为复杂的基因调节和蛋白质产物的功能可用以说明蝇类的更高的复杂性。第二，平行进化同源物的数量在蠕虫和蝇类中显著增加。这反映了蠕虫和蝇类几乎大约一半的基因是其他基因的复制。含有复制基因的基因家族常常形成相同染色体上的基因簇。

最近完成的人类基因组序列测定表明人类基因组有 30 000～40 000 个基因。人类的复杂性与蠕虫相比有极大的不同，人类基因组仅是蠕虫编码基因的两倍，这确实令人感到惊奇。人类基因组中的单一基因与平行进化同源物的数目还不能

确切估计。然而，已经公认的是人类的复杂性在于人类蛋白质组的多样性，而不是人类基因组的大小。

2.6 基因表达、密码子偏倚和蛋白质水平

研究蛋白质组的关键课题之一是细胞中某个特定蛋白质的表达水平。蛋白质表达水平变化极大，从几个拷贝到多于百万个拷贝；但细胞中蛋白质表达水平与其意义大小无关。中间代谢的关键酶或结构蛋白质在每个细胞中有几千拷贝或更多，而与细胞周期调节有关的某些蛋白激酶在每个细胞中仅有几十个拷贝。酿酒酵母有 6 000 个基因，根据 mRNA 水平推测，有大约 4 000 个基因在任何时间都表达。

细胞中某一蛋白质在某一特定时间的表达由下列因素控制：①基因的转录速度；②mRNA 翻译成蛋白质的效率；③细胞中蛋白质的降解速度。基因表达确实在很大程度上决定蛋白质水平。然而，有些研究表明基因表达自身并不与蛋白质水平紧密相关。这一发现也证明了前面提到的 mRNA 的翻译效率和蛋白质降解速度对细胞内蛋白质表达水平的影响，同时也指出了基因表达分析（如微阵）的局限性。

许多基因受可诱导转录因子的调节，这些转录因子又是受多种外界环境影响因素的调节。许多基因表达水平也受内部决定因素——“密码子偏倚”现象的影响。“密码子偏倚”描述了一个生物为相同氨基酸编码时偏向于使用某些密码子。所以，含有不常使用的密码子的基因倾向于较低水平表达。根据酵母基因计算的密码子偏倚值约从 0.2 到 1.0。密码子偏倚值为 1.0 时有最高水平基因表达。大多数酵母基因的密码子偏倚值小于 0.25，表明这些基因在相对低水平表达。

比较酵母某些蛋白质的蛋白质水平、mRNA 表达和密码子偏倚值，尽管在某些方面不完全一致，但可以概括如下。

- 具有低密码子偏倚值的基因倾向于低水平表达，无论根据 mRNA 表达或蛋白质水平的分析都是如此。
- 当基因的密码子偏倚值为 0.25 或更低时（如大多数基因），mRNA 水平与蛋白质水平的相关性很差（$r<0.4$）。对大多数高表达的基因（密码子偏倚值大于 0.5 的基因），mRNA 水平与蛋白质水平的相关性要高很多（$r>0.85$）。
- 长寿蛋白质比短命蛋白质（迅速降解的蛋白质）以明显较高的丰度存在。

因而尽管基因表达测定可能表明蛋白质水平的改变，但很难从基因表达水平推断蛋白质的水平。

2.7 分析蛋白质组学的结论和意义

在任何生物中，蛋白质组是所有基因产物的 30%～80%的汇集。虽然某些蛋白质以较高水平表达（每个细胞 10^4～10^6），但是大多数蛋白质都以相对低的

水平表达（每个细胞 $10^1 \sim 10^2$），与基因表达的绝对量无关，大多数蛋白质以多种翻译后修饰形式存在。这样就给蛋白质组学提出了巨大挑战：我们必须找到测定大量不同蛋白质种类的方法。这些蛋白质大多以相对低水平存在，很多以多种修饰形式存在。本书的下一部分描述可以用来应对这个令人生畏的问题的工具。

推荐读物

Apweiler，R.，Attwood，T. K.，Bairoch，A.，Bateman，A.，Birney，E.，et al.（2001）The InterPro database，an integrated documentation resource for protein families，domains and functional sites. *Nucleic Acids Res*. **29**，37—40.

Coghlan，A. and Wolfe，K. H.（2000）Relationship of codon bias to mRNA concentration and protein length in Saccharomyces cerevisiae. *Yeast* **16**，1131—1145.

Gygi，S. P.，Rochon，Y.，Franza，B. R.，and Aebersold，R.（1999）Correlation between protein and mRNA abundance in yeast. *Mol. Cell Biol*. **19**，1720—1730.

Rubin，G. M.，Yandell，M. D.，Wortman，J. R.，Gabor Miklos，G. L.，Nelson，C. R.，et al.（2000）Comparative genomics of the eukaryotes. *Science* **287**，2204—2215.

Venter，J. C.，Adams，M. D.，Myers，E. W.，Li，P. W.，Mural，R. J.，et al.（2001）The sequence of the human genome. *Science* **291**，1304—1351.



Ⅱ　蛋白质组学的工具



3 分析蛋白质组学概述

在详细讨论分析蛋白质组学之前，我们先概述一些基本方法。蛋白质分析鉴定建立在这样一个基本事实上：大多数含有 6 个或 6 个以上氨基酸的肽序列在一个生物的蛋白质组中是惟一的。换句话说，我们可以将一个 6 氨基酸肽定位于单一基因产物中。因而，如果能得到肽的序列，或者能精确测定肽的质量，就可以通过与蛋白质序列数据库的匹配来鉴定肽片段的蛋白质来源（图 3.1）。当然某些 6 肽可能定位于多个蛋白质，但典型的多次“命中”来自相关蛋白质的高度保守区域（如在第 2 章讨论的平行进化同源物）。如果可以得到定位于相同蛋白质的几个肽序列，这将加强匹配的准确性。因而分析蛋白质组学的本质是将蛋白质转换成肽，得到肽的序列，然后根据在数据库中的序列匹配鉴定相关的蛋白质。

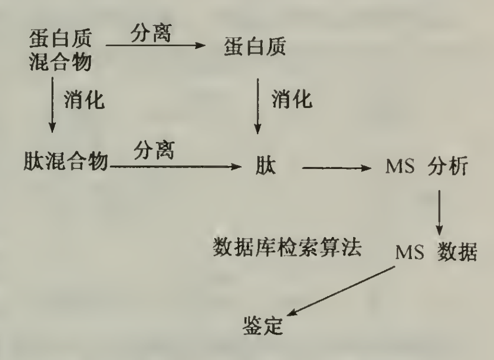


图 3.1　蛋白质组分析的基本流程图

图 3.1 描述了分析蛋白质组学的要素。大多数分析蛋白质组学的问题始于蛋白质混合物。混合物中含有不同分子质量、不同修饰作用和不同溶解度的完整蛋白质。蛋白质必须经切割消化成多肽片段，因为质谱仪往往不能直接对完整的蛋白质进行质量和序列的测定。现代 MS 仪器可以分析复杂的肽混合物，但是组分相对简化的肽混合物更有利于收集数据和分析。

因而，利用 MS 分析蛋白质混合物时，需将含有多种组分的复杂混合物分离，以获得组分较少的简单混合物。先分离出完整的蛋白质，然后消化切割成肽；也可以先将蛋白质切割成肽，在分析前分离肽。第 4 章和第 5 章描述了蛋白质和肽的分离以及将蛋白质切割成肽。

有两种类型的质谱仪可用于分析多肽。第一种类型属于基质辅助激光解吸电离-飞行时间（MALDI-TOF）仪器，主要用以测定多肽的质量。第二种类型属于电喷雾电离（ESI）-串联 MS 仪器，用于分析多肽的序列数据。在第 6 章将对这些仪器进行描述。

在特定软件的帮助下，将质谱仪数据与数据库中的肽序列进行比对以鉴定肽和肽序列。这样可基本上确定混合物中蛋白质的特性。进行这种类型的匹配比对，不需经过直接的 MS 数据分析。第 7 章至第 9 章描述这些软件工具的使用和

蛋白质鉴定方法。

分析蛋白质组学总的来说是一个测定过程。在测定中，蛋白质混合物转换成肽混合物，得到肽 MS 数据，在软件帮助下进行数据库检索，鉴定相关的蛋白质。蛋白质组学之所以作用强大是因为这种测定可用于从各种实验设计产生的多种不同蛋白质样品。蛋白质组学之所以是多用途的是因为通过这种测定可以分析用各种“前端”实验得到的样品。这些前端实验和它们的应用是本书第Ⅲ部分的主要内容。

4　蛋白质和肽的分析分离

4.1　概　　述

这一章描述用于 MS 分析的蛋白质样品的分离方法。在蛋白质组分析的这一阶段必须考虑以下两个方面（图 4.1）。首先，需将完整蛋白质转换成肽。常利用蛋白水解酶对蛋白质进行消化。第二，需将高度复杂的蛋白质和肽混合物分离形成较简单的混合物。这样 MS 仪器能更好地获得混合物各组分的有用数据。这两个步骤没有固定的次序。可以先分离蛋白质，然后消化并分析肽。也可以先将复杂的蛋白质混合物消化成肽，然后分离肽。在这里将讨论每一种方法的优缺点。

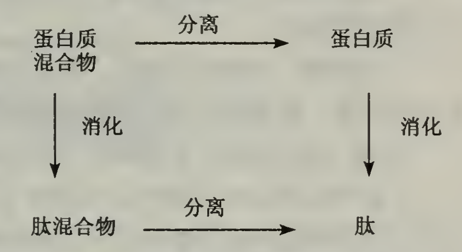


图 4.1　蛋白质组学分析中蛋白质的分离和消化

4.2　复杂蛋白质和肽混合物

在讨论分离和消化方法之前，让我们先考虑复杂蛋白质混合物的问题。MS 仪器能够从相对复杂的混合物样品中获得肽数据。然而，当样品混合物的复杂程度降低时，MS 能鉴定很多的肽片段。样品复杂性和如何处理样品的复杂性可以比作印刷一本书。假如将书中所有的字都印在一张纸上，会产生一张被油墨弄得基本上全黑的纸；而把要印刷的字分配到多张纸上，降低复杂性，可以很容易地阅读每一张纸上的字。对蛋白质和肽的分离我们采用类似的方法，基本上是以“一次一张纸”的方式将肽混合物引入 MS，让仪器尽可能地阅读纸上的内容。

在分离不同类型的蛋白质和肽时，应该先考虑在蛋白质组分析中需要处理的不同蛋白质和肽的数量。根据已知人类基因组的数量，一个人细胞中含有约 20 000个不同表达水平的蛋白质。假定其平均大小为 50 kDa，含有平均水平的赖氨酸和精氨酸，那么每个蛋白质产生约 30 个胰蛋白酶肽。这样一个细胞的蛋白质可产生约 600 000 个胰蛋白酶肽。下文中将看到，这一数字即使对最有效的多维蛋白质和肽分离策略也是一个极大的挑战。

4.3　从生物样品抽提蛋白质

在实际研究中，我们首先是收集生物样品：一块组织、一个平板上培养的细

胞、一瓶细菌、一片叶子等等。样品通常经研磨、匀浆、超声破碎或其他的破裂方法，产生在水缓冲液或悬浮液中含有细胞、亚细胞组分和其他细胞碎片的羹汤般黏稠物质。通过若干技术从这种羹汤般黏稠物中抽提蛋白质。对于蛋白质组分析，目标是回收尽可能多的蛋白质，并尽可能地减少其他生物材料（如脂类、纤维素、核酸等）的污染。抽提蛋白质时常用到以下试剂：

• 去污剂（如 SDS、3-［（3-胆胺丙基）二甲基氨］-1-丙烷磺酸酯(CHAPS)、胆酸盐、吐温)，这些试剂有助于溶解膜蛋白质，并有助于膜蛋白质与脂类的分离。

• 还原剂［如二硫苏糖醇（DTT)、巯基乙醇、硫脲]，用于还原二硫键或防止蛋白质氧化。

• 变性剂（如尿素和酸)，用于改变溶液离子强度和 pH，破坏蛋白质-蛋白质相互作用，破坏蛋白质的二级和三级结构。

• 酶（如 DNase，RNase)，用于消化污染的核酸、糖和脂类。

上面列出的这些试剂可以不同的方式结合使用，生物学各个领域的研究者发展了从各种样品类型（如叶子和培养的细胞）中抽提蛋白质的方法。在某些方法中，通常使用蛋白酶抑制剂防止蛋白酶降解蛋白质。简言之，有多种从生物样品中抽提蛋白质的方法。

必须注意某些试剂可能干扰蛋白质组分析。例如，丝氨酸蛋白酶抑制剂苯甲基磺酰氟（PMSF）常用来在组织加工时防止蛋白质降解。然而，在某些蛋白质样品中残存的 PMSF 可能抑制胰蛋白酶的消化作用。类似地，去污剂可能干扰蛋白质分离和蛋白酶解消化。了解样品的制备过程，对样品分析的成败是很重要的。

4.4 完整蛋白质的分离

广泛使用的，用于完整蛋白质分离的三种主要方法是 1D-SDS-PAGE、2D-SDS-PAGE 和制备等电聚焦（IEF)。还有一些其他方法，特别是 HPLC［反相(RP)、大小排阻、离子交换或亲和层析]，也用于完整蛋白质的分离。分离完整蛋白质是利用了完整蛋白质的物理性质（特别是等电点和分子质量）的不同。样品混合物可以分成较少的组分（如在 1D-SDS-PAGE 和制备 IEF 中)，或分成多种组分（在 2D-SDS-PAGE 中有许多点)。这些组分分别进行蛋白酶消化，然后进一步分离肽片段或直接进行肽 MS 分析。

4.5 1D-SDS-PAGE

这种在蛋白质化学中广泛使用的单向分析分离方法也可用于蛋白质组分析。在 1D-SDS-PAGE 中，蛋白质样品溶解在通常含有巯基还原剂（巯基乙醇或

DTT）和 SDS 的上样缓冲液中（图 4.2）。基本原理是 SDS 与蛋白质结合，以与分子质量约恒定的比例将负电荷（来自 SDS 硫酸基团）传递给蛋白质。高电压下，蛋白质-SDS 复合物在交联的聚丙烯酰胺凝胶上迁移，其速度取决于它们穿越凝胶孔基质的能力。蛋白质按照分子质量次序分离形成条带。

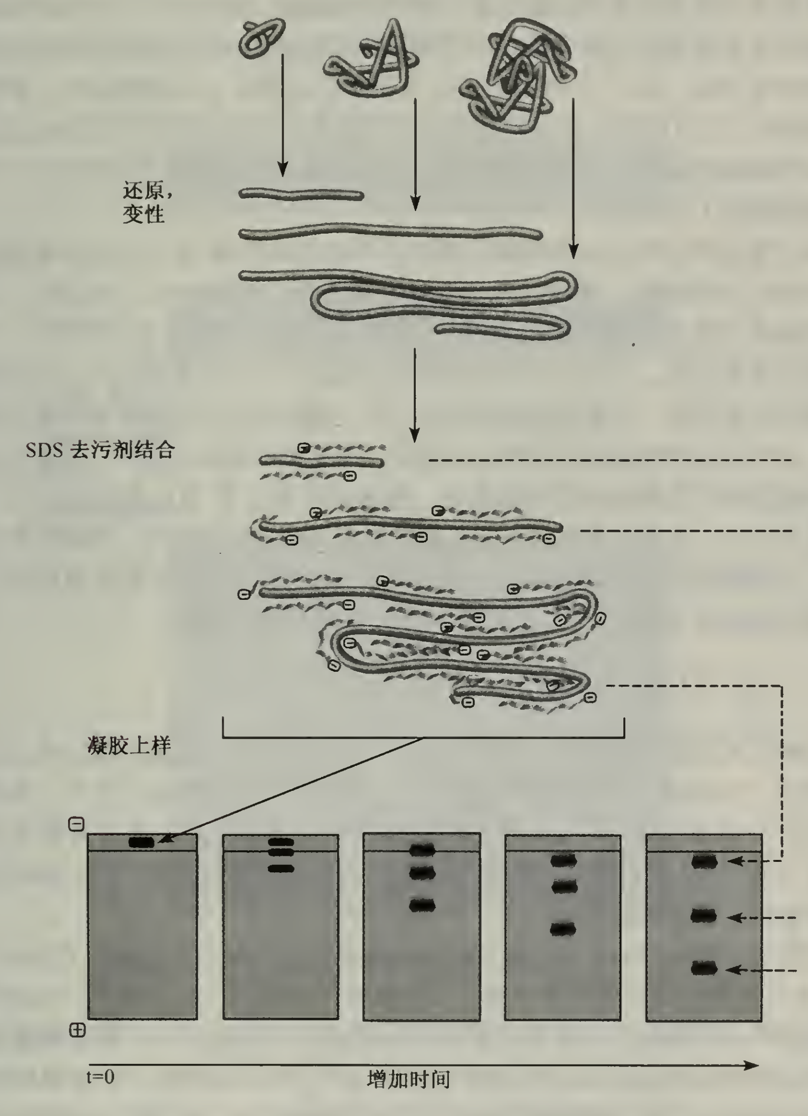


图 4.2　1D-SDS-PAGE

样品的 1D-SDS-PAGE 在交联（即丙烯酰胺的聚合）度为 5%～15%的凝胶上电泳，较大蛋白质在低度交联凝胶中较容易通过。可以根据样品中蛋白质的预测特征来选择交联度。例如，含有低分子质量蛋白质的样品可以在较高交联度的

凝胶上得到较好分离。也可以选择梯度凝胶，交联度从凝胶的顶部到底部逐渐增加。梯度凝胶可对大范围分子质量的蛋白质进行较好的分离。

用 1D-SDS-PAGE 得到的分离度是相当有限的，显示含有单一蛋白质的条带实际上可能含有多种蛋白质分子。细胞粗提物的电泳凝胶上一个跨越约 5 kDa 范围的凝胶切片可能含有几十到几百个不同的蛋白质。甚至一个“纯化的蛋白质”也可能含有多种蛋白质分子形式。比较蛋白质样品的 1D 和 2D-SDS-PAGE 可以很明显的看到这一点。1D-SDS-PAGE 分析常显示出看上去纯净的单一条带，而相同样品的 2D-SDS-PAGE 可将相同分子质量条带分解成具有不同等电点的多个点。这可能反映蛋白质的翻译后修饰，而化学修饰几乎不影响 SDS 结合或在聚丙烯酰胺凝胶上的迁移。

由于蛋白质分离的目的是降低蛋白质混合物的复杂性，据前面提到的 1D-SDS-PAGE 的局限性，使其似乎在蛋白质组分析中用处不大。而实际上这种分离方法的使用取决于样品的复杂性。大多数 1D-SDS-PAGE 在长度为 5～15 cm 的泳道分离蛋白质。这可以很容易将凝胶切割成 5～50 条凝胶切片。对于高度复杂的蛋白质混合物，如完整细胞抽提物，每一凝胶切片可能仍含有许多不同的蛋白质，这样获得的样品仍不够简化。然而，用于蛋白质组分析的许多样品并不是完整细胞抽提物或类似的复杂混合物。例如研究蛋白质-蛋白质相互作用（将在后面几章讨论）的蛋白质组学方法可能含有相对较少的蛋白质。同样，许多生物体液［如脑脊液（CSF）、肺内衬液（lung-lining fluid）］含有较少蛋白质，对于这些混合物的预先分离，1D-SDS-PAGE 可能是相当合适的。

4.6　2D-SDS-PAGE

这种分离方法与蛋白质组学是同义的，一直是分离高度复杂蛋白质混合物的最好方法。2D-SDS-PAGE 实际上是两种不同分离方法的结合。首先，根据等电点用 IEF 分离蛋白质。第二，在聚丙烯酰胺凝胶上电泳进一步分离聚焦的蛋白质（图 4.3）。2D-SDS-PAGE 在第一向电泳根据等电点的不同，第二向电泳根据分子质量的不同分离蛋白质。

尽管 2D-SDS-PAGE 是分离复杂蛋白质混合物的最有效方法，但是在 20 世纪 70 年代早期采用后的许多年来，没有得到广泛使用。这反映了：①进行 IEF 步骤的技术相对困难；②使等电聚焦的蛋白质进入 SDS-PAGE 凝胶的困难。在 2D-SDS-PAGE 的最初形式中，IEF 步骤依赖于“管式凝胶”，这种凝胶需要许多技巧进行操作。而且管式凝胶中的 pH 梯度很难重复。细软管式凝胶中含有的等电聚焦蛋白质很难有效地转移到 SDS-PAGE 平板凝胶上。因而，2D-SDS-PAGE 难于操作，重复性差。

新式 2D-SDS-PAGE 系统的引入大大改善了这一状况。新系统使用固相 pH 梯度（IPG）胶条和相对简易的硬件，能很容易地从 IPG 胶条将蛋白质转

移到 SDS-PAGE 平板凝胶中。IPG 胶条使用固相 pH 梯度。在固相 pH 梯度中，聚羧酸两性电解质固定在支持物上，产生可重复的稳定 pH 梯度。现在可以从主要供应商购买在各种或宽或窄 pH 范围重复分离的 IPG 胶条。使用窄 pH 范围有助于分离具有相似等电点的蛋白质。图 4.3 总结了 IEF 分离步骤。胶条用缓冲液水合，蛋白质在电压下缓慢加样到胶条。然后增加电压得以聚焦。商业上已有的系统能够提供温度控制以及高精度电压或电流控制以便进行可重复的分离。

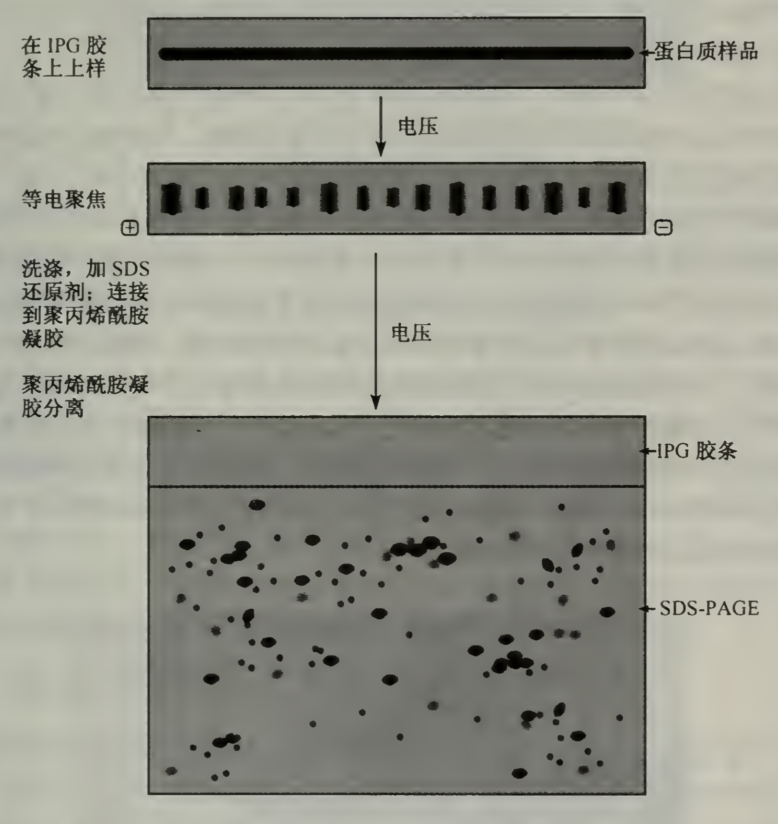


图 4.3　2D-SDS-PAGE

在聚焦步骤后，胶条用含有巯基还原剂和 SDS 的缓冲液处理，然后连接到 SDS-PAGE 平板凝胶中。在这方面，含有聚焦的蛋白质的 IPG 胶条类似于 1D-SDS-PAGE 的“浓缩胶”。蛋白质在 SDS-PAGE 平板凝胶中以与 1D-SDS-PAGE 相同的方式进行分离。通过 2D 凝胶分离蛋白质的显色使用通常的染色技术，包括银染、考马斯亮蓝和酰胺黑染色。银染和较新的荧光染料是最灵敏的。尽管可采用许多不同的方法进行染色，但不是所有的方法都适用于其后的蛋白质分析。例如，用甲醛固定的蛋白质银染倾向固定胶中的蛋白质，阻止蛋白质的消化和所

形成肽的回收。凝胶长时间暴露在乙酸中也引起类似的问题。因而，使用适合于消化和洗脱步骤的染色方法是重要的。

4.7 2D-SDS-PAGE 存在的问题

尽管作为分离复杂蛋白质混合物的方法，2D-SDS-PAGE 相对其他方法有许多优势，但这项技术也有某些问题。首先是进行可完全重复的 2D-SDS-PAGE 分析是困难的。通过比较 2D-SDS-PAGE 染色凝胶的图像来比较两个样品时，这个问题就凸显出来。一维蛋白质迁移的任何不同可能会被误认为是两个样品中某些蛋白质的不同。

2D-SDS-PAGE 的第二个问题是某些蛋白质不太适于进行第一维 IEF 步骤。许多较大的疏水蛋白质在这种类型的分析中结果不理想。这些蛋白质的低溶解度导致蛋白质沉淀和聚集，造成在 IPG 胶条中蛋白质的“成片条带”，而不是形成不连续条带的清晰聚焦。当这些蛋白质再进行第二维（SDS-PAGE）电泳时，显示为横越分子质量区域的条纹（图 4.4）。还有另一个相关问题，它或者是这种电泳 2D-SDS-PAGE 的优点，或者是缺点。由于蛋白质存在不同等电点的多种形式，蛋白质的 IEF 常把蛋白质分离成许多不连续条带。例如，使中性酰胺转换成阴离子羧基的脱酰胺化可以改变蛋白质的 pI 和它在 IPG 胶条的迁移。其他可以影响 pI 的修饰包括糖基化、磷酸化、氧化和外源化学修饰。在某些情况下，相同多肽的不同修饰可能显现为“点横向排列”（图 4.5）。尽管这种分离对建立蛋白质不同的存在形式是有用的，但它也可能使通过 2D-SDS-PAGE 估计两个样品中相关蛋白质表达的问题更复杂。



图 4.4 在 IEF（水平）方向显示蛋白质“成片条带”的部分 2D 凝胶



图 4.5 由于相同蛋白质的不同修饰/电荷显示
“点横向排列”的部分 2D 凝胶

2D-SDS-PAGE 的第三个问题是作为检测技术的蛋白质染色的动态范围相对较小。点密度最多能反映 100 倍的蛋白质浓度范围。这意味着 2D 凝胶的染色能看到高含量的蛋白质，而较低含量的蛋白质往往不能检测。一个极好的例子来自 Steven Gygi 和 Ruedi Abersold 的工作，他们研究了酵母基因表达（通过测定 mRNA）和蛋白质水平（通过测定同位素标记的甲硫氨酸的掺入）的关系。酵母表达其 6 000 个基因的三分之二，然而用银染色的 2D-SDS-PAGE 的精细分析最多能检测约 1 000 个蛋白质。换言之，在约 4 000 个表达的基因中，有 3 000 个在 2D-SDS-PAGE 中不能检测。大多数检测到的蛋白质具有高密码子偏倚值的基因（见第 2 章），倾向于较高水平的表达。2D-SDS-PAGE 最适宜用来分析高含量的长寿蛋白质。不幸的是许多生物学上非常重要的蛋白质以相对低水平表达，并迅速转换，因此引入其他分析方法是必要的。

4.8 制备 IEF

这项技术类似于 2D-SDS-PAGE 的第一步。在制备 IEF 中，使用 IPG 胶条，在一个管胶中或在溶液中进行分离。后者是最常使用的。用可溶性两性电解质得到 pH 梯度。两性电解质是聚羧酸化合物，当电压加到聚焦池时产生稳定 pH 梯度。然后加入蛋白质样品，加电压，蛋白质通过等电点进行分离。商业上已有的装置，如 BioRad Rotofor™ cell，聚焦池由可渗透膜分成一系列小室。在聚焦后，用真空吸入器（vacuum sipper）同时迅速的把小室抽空。真空吸入器把聚焦池每一部分的物质吸入一个试管。用这种类型装置，把整个蛋白质混合物分成 12～20个组分。

溶液等电聚焦的优点是相对较大的样品容量（每一次可使用毫克到克的总蛋白质）和样品在溶液中比在凝胶中更容易操作。在进一步处理蛋白质前，可通过透析或凝胶过滤把两性电解质从分部的样品中除去。蛋白质从液相 IEF 的回收

率大于 85%～90%。可以用去污剂和促溶剂保持疏水蛋白质的可溶性。和 2D-SDS-PAGE 的 IEF 步骤一样，这种分离利用完整蛋白质物理性质（pI）的多样性。操作多种完整蛋白质也有缺点，如在溶液相聚焦时，某些蛋白质易发生聚集和沉淀。

4.9 高效液相层析

改良的固定相材料和硬件的出现极大地改进了用于蛋白质纯化的 LC 操作系统。对分析蛋白质组学来说，尽管完整蛋白质的 HPLC 还没有成为广泛使用的技术，但 HPLC 仍可作为分离蛋白质混合物的最初步骤。HPLC 可进行各种层析分离，包括 RP、阴离子和阳离子交换、大小排阻和亲和层析。亲和层析可以从复杂混合物中分离某种蛋白质组分，这是其很大的优点。

HPLC 能把蛋白质混合物分离成各种组分，这与制备 IEF 的功能相同。HPLC 的优点是有各种分离模式。串联 HPLC 结合了两种不同类型的层析，例如强阳离子交换之后连接 RP，使用两种完全不同的分离模式。正如下面将讨论的有关肽的 HPLC，离子交换可与 RP 连接获取高效率串联 LC 分离。

4.10 肽分离

在这种方法中，样品中的蛋白质首先消化成肽混合物，然后进行肽分离。这种方法应用的极致是把全部细胞或组织抽提物消化成肽，然后进行肽混合物的 MS 分析。这类分析已取得了极大成功。具有特殊控制的微毛细管 HPLC 和自动控制 MS 仪器的使用（本书后面讨论），可以一次获取几百或几千个肽的 MS 数据。这种方法的基本原理是将非常不均一的蛋白质混合物转换成更容易分析的较均一的肽混合物。如果确实选用这个方法，分离肽混合物的方法是很少的。1D-SDS-PAGE 和 2D-SDS-PAGE 被排除，因为它们实际上不能用于从消化物中分离肽。消化产物的 pI 和分子质量范围相差很小。尽管制备 IEF 可以用于肽混合物分离，但在分离肽混合物方面的用途可能有限。然而，评估制备 IEF 作为肽分离工具的工作做得很少，它的作用不能完全排除。

4.11 肽分析的串联 LC 方法

常用的分析肽混合物的方法是 HPLC。如前所述，固定相和分离模式的多样性使 HPLC 有很大的分离能力。HPLC 分离模式的结合使用是分析蛋白质组学中最有效的工具之一。连续使用各种不同的分离模式称为“串联 HPLC”。串联 LC 原理是不同分离模式的结合使用可使混合物中的肽得到更大程度的分离。考虑以下 HPLC 主要分离模式和特征。

- RP：疏水性。
- 强阳离子交换：净正电荷。

• 强阴离子交换：净负电荷。
• 大小排阻：肽大小/分子质量。
• 亲和：与特定功能基团的相互作用。

在以上列出的分离模式中，除大小排阻以外其余都能用于肽分离。已有的大小排阻介质的分离能力不能分离分子质量相近的肽。

John Yates 和同事开发出分析复杂肽混合物的串联 LC-MS。其方法使用连续连接的毛细管柱，直接洗脱进入质谱仪（图 4.6）。他们创造了“MudPIT”一词（多维蛋白质鉴定技术）来描述这种方法。肽先用强阳离子交换（SCX）柱分离。SCX 柱作为分析系统的“前端”（图 4.7）。肽吸附到具有亲和力的 SCX 柱，亲和力与每一个肽的总正电荷（如离子化的氮）数目成比例。通过不断增加的盐浓度阶梯梯度洗脱肽。每一梯度释放一组肽。这组肽再传递到 SCX 柱下游的 RP 柱上，通过 RP-HPLC 梯度分离。RP-HPLC 梯度根据疏水性分离肽。肽直接从 RP 柱传递进入 MS 仪器供进一步分析。在 RP 梯度分离完成后，另一组从 SCX 柱释放的肽再经 RP 柱分离，直接进入 MS 进行分析。这种循环一直持续到所有的肽从 SCX 柱中洗脱完成。

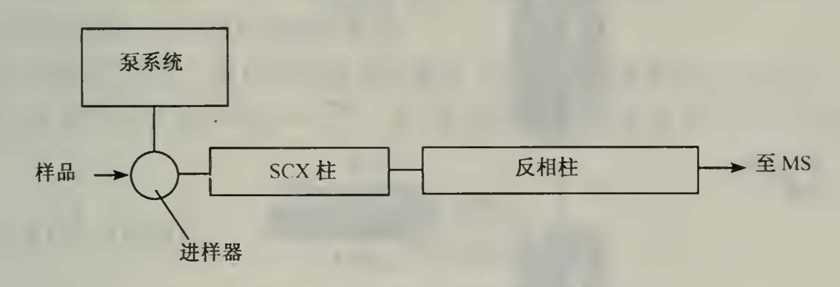


图 4.6　串联离子交换（SCX）-RP-HPLC 系统

比较 MudPIT 串联 LC 方法和单独用于 LC-MS 的 RP-HPLC，很明显串联方法大大增加了所鉴定肽的数量。串联方法用于肽混合物的进一步“分散”，以便 MS 可从更多更细的组分分离中获取数据。此外串联 LC 方法也有助于鉴定来自混合物中低丰度蛋白质的肽。用 MudPIT 方法分析酵母蛋白质揭示这种方法可以更好地鉴定低丰度蛋白质的酶消化肽。这与适于鉴定较高表达蛋白质的 2D-SDS-PAGE 明显不同。

串联 LC 比 2D-SDS-PAGE 具有更多的优势，主要是因为以下两点：首先，2D 凝胶是对染色显示的蛋白质点进行消化和 MS 分析，而许多 MS 仪器测定的范围低于凝胶上染色显示的水平。在 2D-SDS-PAGE 上，如果染色不能显示需收集和分析的蛋白质点，就收集不到此蛋白质的数据。第二，混合物中蛋白质的处理可能提供“载体效应”，即较高丰度肽的存在可防止低丰度肽的丢失。来自相对较弱的 2D 凝胶蛋白质点的样品浓度很低，由于与仪器表面的相互作用等造

成样品有比例较大的一部分丢失，而串联 LC 系统分析的是复杂混合物中的多肽，在与仪器表面相互作用时由于其他大量组分的存在，会降低低丰度组分的丢失。

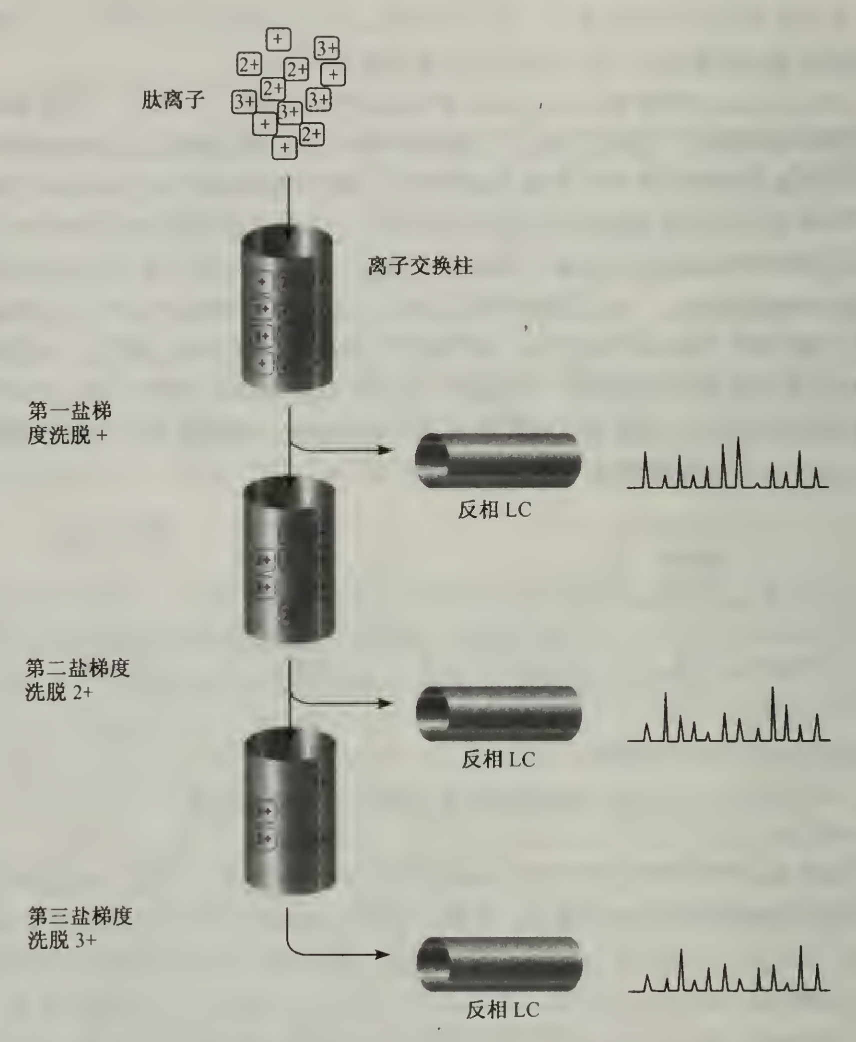


图 4.7 肽混合物用线性强阳离子交换和 RP-HPLC 进行的多步分离

也可以采用其他 LC 分离模式的结合，包括强阴离子交换/RP-HPLC 和亲和/RP-HPLC 的结合使用。对蛋白质组分析来说，串联 LC 技术仍是新技术，其应用前景会加速其发展，最终将成为广泛使用的方法。

4.12 毛细管电泳

毛细管电泳（CE）与 IEF 的基本原理相同。置于电场中的蛋白质迁移到一

个 pH 梯度点。在这个 pH 梯度点蛋白质净电荷为中性。在微毛细管中的分析操作比前面讨论的制备 IEF 技术有很大提高。在所有肽分析技术中，CE 具有最高的分离能力，可以直接与 MS 仪器连接。CE 作为一项分析蛋白质组学的技术具有极大的潜力。目前由于缺乏用于分析蛋白质组学的商业上稳定可靠的 CE-MS 仪器，使得 CE 的应用受到限制。这方面仪器开发正在继续，很快 CE-MS 将成为蛋白质组学分析中非常有用的工具。

4.13 哪种方法最好？

先进行蛋白质分离，然后进行消化和分析是当今最常用的分析蛋白质组学方法。这很大程度上依赖于进行蛋白质分离的 2D-SDS-PAGE。这一方法的一个最大优点是 2D 凝胶可作为图像图谱，研究人员可根据凝胶上点的图形变化比较蛋白质组的变化。如前所述，虽然仍有一些因素影响 2D 凝胶点图像的解释，然而其他技术都不能进行蛋白质组的直观“快照”。所以，2D-SDS-PAGE 可能继续作为蛋白质组学的主要方法学。但对于低丰度蛋白质，2D 凝胶发挥的作用有限，因为不能从凝胶上看到一些重要的低丰度蛋白质。在这种情况下，其他的分离方法，特别是串联 LC 是很好的替代技术。

鉴于前面的讨论，蛋白质组鉴定的最好方法可能是各种方法的混合。一种常见的混合使用方法见图4.8。第一步，通过制备IEF、制备1D-SDS-PAGE或

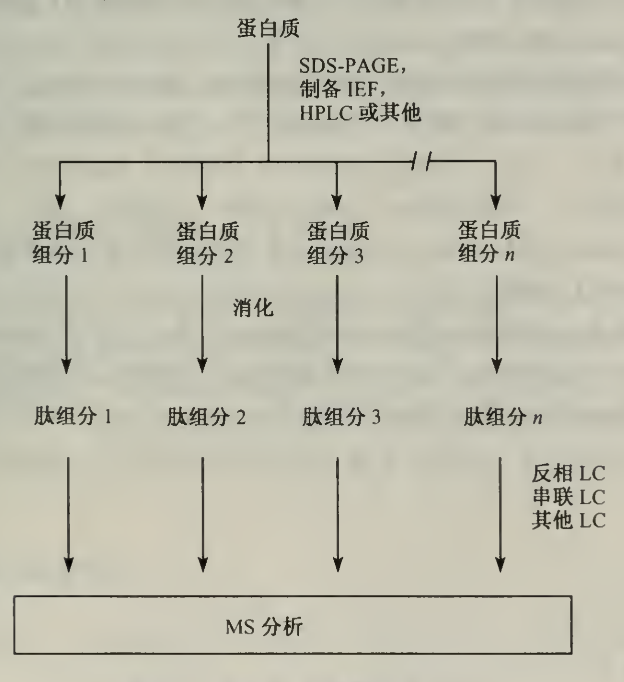


图 4.8 蛋白质/肽分离的基本方法

HPLC 分离完整蛋白质，然后对分离得到的组分进行酶消化，所产生的肽在进入 MS 之前进行 HPLC 分离。由于样品的复杂性或分析目的的不同，HPLC 分离可以是一个单分离模式（如 RP）或是一个串联 LC 分离模式。这种方法的一个突出优点是其灵活性和易于适合不同实验室的仪器操作。这种方法的另一个优点是前端蛋白质分离可以处理量较大的蛋白质（在大多数情况下可处理若干毫克），这样提高了 MS 分析中检测低丰度肽组分的可能性。最近发表了其他几种可用的分析方法，但仍然需要有进一步的工作明确建立哪些方法是最容易、最有效和最可靠的。

推荐读物

Gygi，S. P.，Corthnls，G. L.，Zhang，Y.，Rochon，Y. and Aebersold，R.（2000）Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Nat. Acad. Sci.* **97**，9390－9395.

Link，A. J.（1998）*2-D Proteome Analysis Protocols*. Humana Press，Totowa，NJ.

Link，A. J.，Eng，J.，Schieltz，D. M.，Carmack，E.，Mize，G. J.，Morris，D. R.，et al.（1999）Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* **17**，676－682.

Rabilloud，T.（2000）Detecting proteins separated by 2D gel electrophoresis. *Anal. Chem.* **72**，48A－55A.

Walker，J. M.（1996）*Protein Protocols Handbook*. Humana Press，Totowa，NJ.

Wall，D. B.，Kachman，M. T.，Gong，S.，Hinderer，R.，Parus，S.，misek，D. E.，et al.（2000）Isoelectric focusing nonporous RP HPLC：a two-dimensional liquid-phase separation method for mapping of cellular proteins with identification using MALDI-TOF mass spectrometry. *Anal. Chem.* **72**，1099－1111.

Washburn，M. P.，Wolters，D.，and Yates，J. R.（2001）Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**，242－247.

5 蛋白质消化技术

5.1 为什么需要消化蛋白质?

现代 MS 仪器能高精度地测定完整蛋白质分子质量。为什么蛋白质组学不简单地测定完整蛋白质的质量? MS 仪器较少对完整蛋白质进行质量测定主要有三个原因。第一，尽管 MS 仪器已经相当完善，但是仪器所进行的测定仍然会发生错误。蛋白质的质量越大，绝对误差值越大。这些误差会引入很高的不确定性，使得鉴定不够准确。而且，各种翻译后修饰使质量测定更为复杂。第二，不是所有的蛋白质都能进行完整蛋白质的质量测定。MS 对大分子蛋白质和疏水蛋白质很难进行质量测定。第三，完整蛋白质质量测定的灵敏度远低于肽质量测定和肽串联 MS 分析的灵敏度。由于这些原因，通过分析完整蛋白质来研究蛋白质组学目前是不现实的。

进行肽分析有较好的方法也是不选择对完整蛋白质进行分析的另一个原因。MS 仪器目前非常适合分析肽。如在下一章将看到的，现代 MS 仪器可以进行肽的高精度质量测定，也能得到序列的数据，而且从 MS 分析得到的肽段数据可直接与蛋白质和核苷酸序列数据库中的蛋白质序列进行比较。通过肽 MS 数据与数据库信息的比较以鉴定蛋白质，这一检索算法的重要基础是某些蛋白水解酶在特定位点将蛋白质切割成肽。在这一章将讨论可切割蛋白质产生肽供 MS 分析所用的酶及其方法。

5.2 消化可以达到什么目的?

理想的消化方法是在某些特定的氨基酸残基上切割蛋白质，产生最适于 MS 分析的片段。6～20 氨基酸的肽片段对 MS 分析和数据库比较最为理想。一般短于 6 氨基酸的肽太短，不能在数据库检索中产生惟一的序列匹配。另一方面，在串联 MS 分析中（下一章将有较详细讨论），很难从长于 20 个氨基酸的肽段上获得序列信息。因而蛋白质消化的目标是尽可能多的获得合适长度的肽片段供 MS 分析。

5.3 蛋白酶概述

大自然进化产生各种蛋白酶以便不断对高等生物进行必要的蛋白质重建。已纯化和鉴定了上千种不同的蛋白酶，但是大多数酶在生物体中的含量很低。生物

表 5.1 蛋白酶及其切割专一性

酶	切割专一性
胰蛋白酶	/K-,/R-,\P
胰凝乳蛋白酶	/W-, /Y-,/F-,\P
Glu-C(V8 蛋白酶)	/E-,/D-[a],\P
LysC	/K-,\P
AspN	/D-

a 在磷酸钠缓冲液中切割天冬氨酸和谷氨酸，在其他情况下只切割谷氨酸。

化学家只能获得可以纯化或表达的蛋白酶。分析蛋白质组学真正需要的是稳定的、有确定专一性并得到很好鉴定的酶。这些酶必须可大量得到、纯度高并且稳定性好能在各种情况下使用。有些符合这些要求的蛋白酶已用于蛋白质组分析。表 5.1 总结了在蛋白质组分析中广泛使用的蛋白酶和它们的切割特性。下面给出几种酶主要特征的小结。

5.4 胰蛋白酶

胰蛋白酶是蛋白质组学分析中最常用的蛋白酶。胰蛋白酶主要从猪或牛的胰脏中获得，易于纯化。胰蛋白酶被对甲苯磺酰基苯丙氨酰氯甲烷（TCPK）修饰，抑制残余的胰凝乳蛋白酶。胰蛋白酶在赖氨酸和精氨酸残基位点切割蛋白质，但若在赖氨酸或精氨酸的 C 端方向有脯氨酸则不能切割。赖氨酸和精氨酸在许多蛋白质中的距离能产生长度很适于 MS 分析的肽。这种“双专一性”意味着胰蛋白酶比只识别一个氨基酸残基的蛋白酶更频繁地切割蛋白质。一般规律是一个 50 kDa 蛋白质产生约 30 个胰蛋白酶肽。

对蛋白质组学工作来说，胰蛋白酶的一个优点是该酶在溶液和“在凝胶中”进行消化（见下文）都有很好的活性。已经发展了若干胰蛋白酶在溶液中、凝胶中或膜印迹上消化蛋白质的方法，这些方法已被广泛测试。经常进行蛋白质组学分析的实验室熟悉胰蛋白酶自溶片段，它作为胰蛋白酶消化方法的副产物是不可避免的。

5.5 Glu-C（V8 蛋白酶）

Glu-C 是内切蛋白酶，它在醋酸氨或碳酸氢氨缓冲液中切割谷氨酸残基的羧基侧，但在磷酸钠缓冲液中切割谷氨酸和天冬氨酸残基。Glu-C 的一个优点是与胰蛋白酶相比有明显不同的切割专一性，这提高了得到蛋白质互补肽片段的可能性。这对分析具有高赖氨酸和精氨酸区域的蛋白质可能特别有用，这些区域能被胰蛋白酶充分切割产生没有足够序列信息的极短的肽片段。

5.6 其他蛋白酶和切割试剂

还有几种酶也可用于蛋白质组分析。这些酶包括 LysC、胰凝乳蛋白酶、Asp-N 和几种“非专一性”蛋白酶。这些酶的切割专一性对大多数蛋白质组分析不理想。这些只在一个氨基酸位点切割的酶产生几个大片段。这些大片段在串联 MS 分析中不能提供有用的序列信息。另一方面，胰凝乳蛋白酶切割太频繁

（能切割酪氨酸、苯丙氨酸和色氨酸位点），产生过多序列太短的小肽。不过，当一个感兴趣的蛋白质序列，特别是在某些感兴趣的区域，不能产生令人满意的胰蛋白酶肽时，可以选择使用这些酶。

5.7 非专一性蛋白酶

非专一性蛋白酶也是蛋白质消化中非常有用的酶，如枯草杆菌细胞溶素蛋白酶（subtilysin）、胃蛋白酶、蛋白酶 K 和链霉蛋白酶。这些酶或多或少地随机切割蛋白质产生多个重叠肽。由于专一性较差，消化必须在相对较短的时间内进行以防止消化得太过，产生多个重叠肽的优点是能获得较多的蛋白质序列数据。

5.8 溴化氰

蛋白质也可用某些化学试剂切割。最常用的化学试剂是溴化氰（CNBr），它在甲硫氨酸位点切割蛋白质。CNBr 的反应具有高度专一性，但是在大多数蛋白质中甲硫氨酸残基含量较少，CNBr 切割产生数量较少的大片段，这些大片段在串联 MS 分析中不能产生有用序列。

5.9 在凝胶中消化蛋白质

1D 或 2D-SDS-PAGE 分离蛋白质的消化通常使用的方法为“在凝胶中”消化（图 5.1）。从凝胶中切出感兴趣的条带或点，进行脱色，然后用蛋白酶（最常用的是胰蛋白酶）处理。酶进入凝胶基质，将蛋白质消化成肽，然后通过漂洗将肽从凝胶中洗脱。这项技术在 2D-SDS-PAGE 蛋白质组学方法中必不可少。

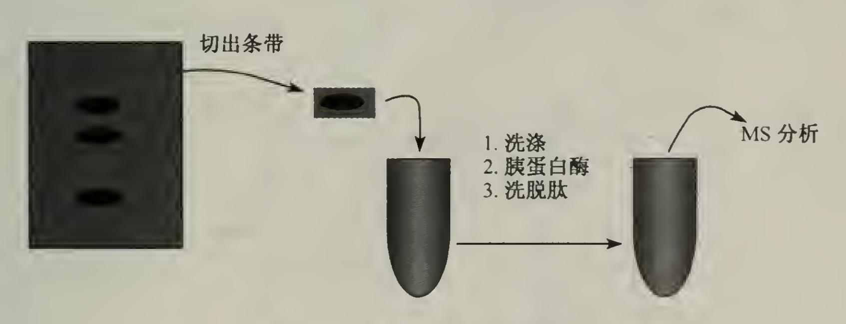


图 5.1 在凝胶中消化蛋白质

胰蛋白酶是最常使用的酶，其他的蛋白酶（包括 Glu-C 和胰凝乳蛋白酶）也可用于这种基本方法。凝胶中蛋白质的消化和肽片段的回收率是受许多因素影响的。在凝胶中能否进行成功消化的一个决定性因素是凝胶染色技术。使用醛固定和长时间暴露在酸（如醋酸）中的染色方法会将蛋白质固定在凝胶中，使蛋白质

不易消化，肽很难洗脱。高度交联的凝胶阻碍蛋白酶进入凝胶基质。最后，SDS-PAGE 技术中残留的组分（如 SDS 或残留的未聚合的丙烯酰胺）对蛋白酶活性可能有抑制作用。印迹转移到硝酸纤维素膜或聚偏氟乙烯（PVDF）膜的蛋白质也可以进行"在膜上"的消化。与在凝胶中消化类似，切出含有感兴趣蛋白质的膜部分，用蛋白酶进行消化，然后从膜表面洗脱肽片段。

推荐读物

Jensen，O. N.，Wilm，M.，Shevchenko，A.，and Mann，M.（1999）Sample preparation methods for mass spectrometric peptide mapping directly from 2-DE gels. *Methods Mol. Biol.* **112**，513—530.

Shevchenko，A.，Wilm，M.，Vorm，O.，and Mann，M.（1996）Mass spectrometric sequencing of proteins silver-stained polyacrylamide gels. *Anal. Chem.* **68**，850—858.

Walker，J. M.（1996）*Protein Protocols Handbook*. Humana press，Totowa，NJ.

6　分析蛋白质和肽的质谱仪

6.1　引言

用于大多数蛋白质组学 MS 分析的仪器有两种类型：MALDI-TOF 仪器和 ESI 串联 MS 仪器。这两种类型的仪器以完全不同的方式工作，产生不同的但是互补的信息。装备最好的蛋白质组学实验室应该具备这两种仪器。这一章将描述每一种仪器的工作过程，产生的数据类型，比较其优点和局限性。在讨论仪器之前，先看一下 MS 仪器操作的原理。

6.2　MS 仪器怎样工作

质谱仪有三个基本部分（图 6.1）。第一部分是离子源，由样品产生离子。第二部分是质量分析器，根据离子的质量/电荷（m/z）比分离离子。第三部分是检测器，检测由质量分析器分离的离子。简言之，质谱仪将混合物的组分转换成离子，然后根据离子的 m/z 进行分析。数据由数据系统自动记录，供研究者进行手工或计算机辅助的解释。

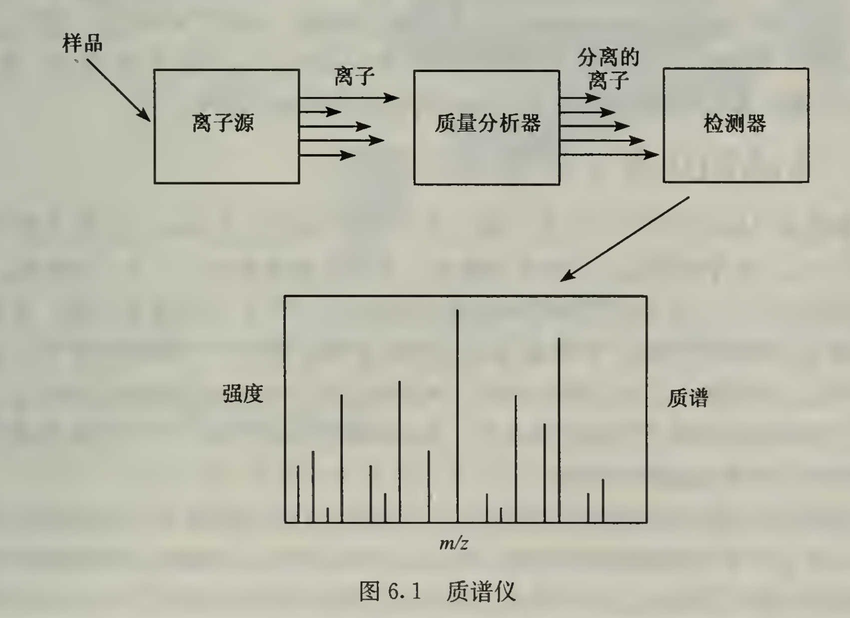


图 6.1　质谱仪

当然，运行 MS 系统还需要更多其他的部分。现代 MS 仪器由复杂的计算机控制，仪器产生的数据由同样复杂的计算机数据系统处理。仪器装备有真空泵系

统，保证质量分析器和检测器处于高真空，这是其行使功能所必须的。与早期的MS仪器相比，现在的MS仪器体积相对较小、设计更紧凑、数据更可靠，更易于操作。

6.3 我们希望从MS数据得到什么？

蛋白质组学的研究需要有肽质量（MALDI-TOF MS）的良好数据和描述肽裂解（ESI串联MS）的良好数据。怎样得到良好数据？我们必须注意三点。第一是灵敏度。如前所述，在大多数蛋白质组学工作中，蛋白质样品的量是有限的。我们需要能分析飞摩尔（10^{-15}摩尔）或样品量更低的仪器。第二是分辨率，它是能在多大程度上区分 m/z 相似离子的测量。较高分辨率的MS仪器（磁分区仪器或傅里叶变换仪器）能够可靠区分 m/z 仅相差0.001 amu（原子质量单位）的离子。这些昂贵、难操作的仪器在蛋白质组学工作中不常使用。在MS中通常使用的仪器能至少区分 m/z 值有1 Da（单个氢原子的质量）不同的离子，某些质量分析器能提供较高的分辨率，在特殊情况下需要较高的分辨率，这将在后面讨论。第三是质量精确性。肽离子或肽片段离子的测定值必须尽可能接近其真实值。这在使用MS数据与（真实）数据库数据比较鉴定肽时特别重要。

6.4 MALDI-TOF MS仪器

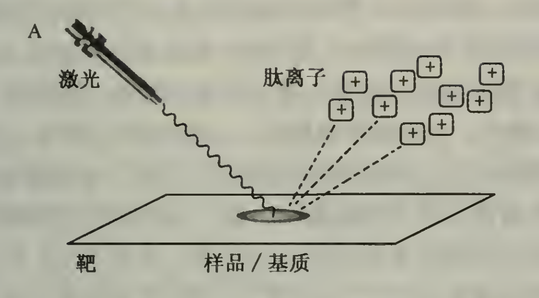
MALDI-TOF是基质辅助激光解吸电离飞行时间质谱的标准缩写。第一部分（MALDI）指离子源，而TOF是指质量分析器。术语“MALDI”实际上描述电离的一种方法，在蛋白质组学文献中常作为MALDI-TOF的速记。然而，MALDI离子源和TOF分析器也可以在其他仪器结构中使用。

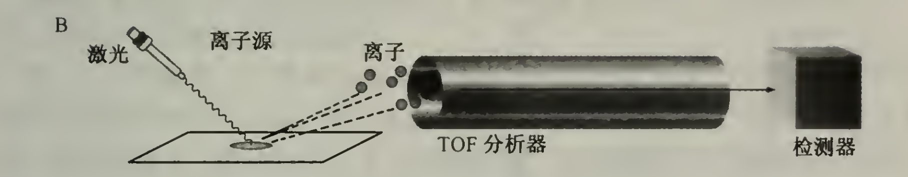
6.5 MALDI离子源怎样工作

要理解MALDI-TOF仪器怎样工作，最简单的是从MALDI离子源开始(图6.2A)。待分析样品与化学基质混合。常用的基质含有一小分子有机物，具有吸收特定波长的生色团。典型的基质化合物包括2，5-二羟基苯甲酸、芥子酸和α-氰基-4-羟基肉桂酸。将样品和基质的混合物点到一个小板或玻片上，在空气中气化，同时将样品分子带入气相。离子源装有激光，向混合物发射光束，基质化合物吸收射线光子形成激发电子。多余的能量转移到样品中的肽片段或蛋白质，使它们从靶表面射入气相。

这种电离过程产生正离子和负离子，这取决于样品的性质。对于肽和蛋白质，正离子常常是我们感兴趣的离子种类。正离子是由于肽或蛋白质从基质发射时接受一个质子而形成。每个肽分子倾向于获取单个质子。因而产生的大多数肽离子带单电荷。对于一个质量为1 032 Da的肽，接受一个质子并带上正电荷，使 $[M+H]^+$ 离子的 m/z 值为1 033。然后提取在MALDI离子源形成的离子并

使离子进入 TOF 质量分析器。





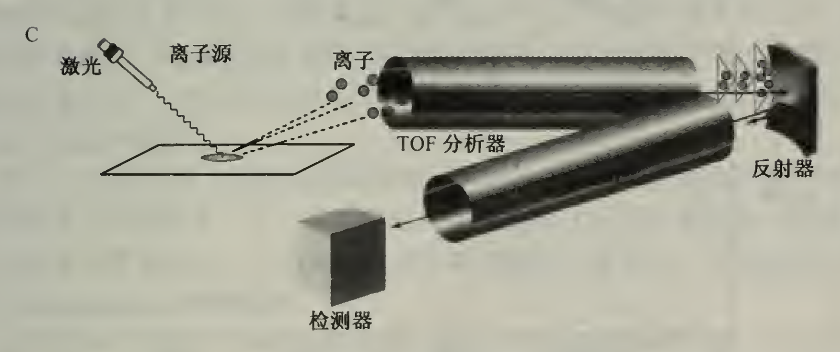


图 6.2　MALDI-TOF 质谱仪的图示

A. MALDI 电离过程；B. 线性模式操作的 MALDI-TOF 仪器；C. 装备有反射器的 MALDI-TOF 仪器。

6.6　TOF 质量分析器

TOF（飞行时间）质量分析器正如其名字所表示的含义。TOF 分析器测定离子从分析器的一端飞行到另一端并撞击检测器所用的时间。离子沿飞行管向下飞行的速度与它们的 m/z 值成反比例。m/z 越大，飞行得越慢。

第一个 TOF 分析器工作方式简单（图 6.2B）。这种简单的“开始到结束”分析器的工作方式称为“线性模式”。在 MALDI 离子源形成离子，连续从离子源提取离子，沿飞行管飞行到检测器。但是用连续提取离子的线性模式工作的

TOF 仪器分辨率相对不高。质谱的分辨率是指仪器区分 m/z 值相近的离子的能力。MS 分辨率可比做是视力聚焦，那么低分辨率就如同近视。线性模式仪器分辨率不高是由于有相同 m/z 的离子沿飞行管向下飞行时，其速度发生变化。

两个重要的技术发明解决了低分辨率的问题。第一个是反射器。按照视力比拟，它相当于近视 TOF 的接触镜片。反射器使有相同 m/z 值的离子聚焦，并使它们在相同时间到达检测器（图 6. 2C）。反射器极大提高了 TOF 分析的分辨率。图 6. 3 的谱图中生动地表明了反射器对分辨率的影响。图 6. 3B 是在线性模式中得到的胰岛素的谱图，表明所分析的胰岛素肽的平均 m/z 值。在图 6. 3A 中，有反射器仪器的分析很容易分辨胰岛素肽所有 ^{12}C 和各种 ^{13}C 同位素体的单个离子。

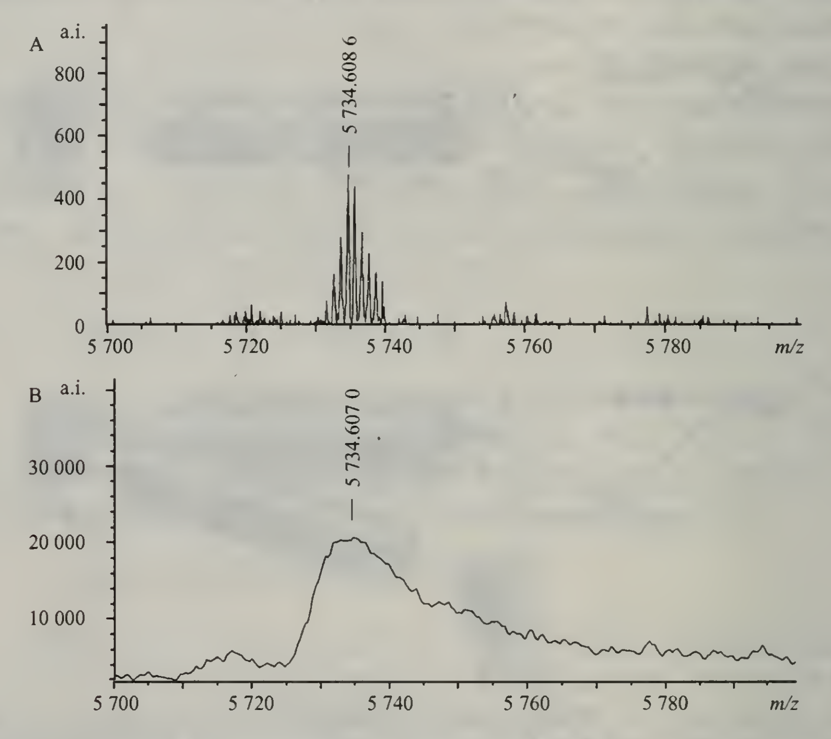


图 6. 3　用反射器模式（A）和线性模式（B）进行的胰岛素的 MALDI-TOF MS 分析

另一个提高线性模式中 TOF 分析器分辨率的方法是使用延迟提取的脉冲激光电离。延迟提取技术在激光脉冲（电离）和沿飞行管的离子方向之间建立一个微小延迟。这使所有离子有一个“公平起始”，从而使所有具有相同 m/z 的离子种类在相同时间撞击检测器。对许多激光脉冲（10～100）获得的谱图进行平均来得到谱图。与反射器类似延迟提取也可以提高谱图分辨率。

TOF 分析器技术的开发产生了某些当前最好的质量分析器。最好的 TOF 分析器的分辨率可以可靠区分 m/z 有 0. 001 amu 不同的肽离子。下一章将看到，在蛋白质鉴定中，高分辨率和质量精确性对 MALDI-TOF 数据的可靠应用是必需的。

在蛋白质组学中使用的 MALDI-TOF 仪器主要是获取完整肽离子的质量测定，某些仪器也能分析肽离子片段。一种称为“源后衰变”（PSD）的技术可用于装备有反射器的 TOF 分析器。PSD 调节反射器的电压，以便能探测肽的裂解和沿飞行管加速时形成的肽离子片段，PSD 对完整肽质量的测定是一个有用的补充技术。肽 PSD 谱图中出现具有基本公式 $H_2N^+=CHR$（R 是氨基酸侧链）的肽亚氨离子，这些亚氨离子是存在特定氨基酸的标志，可在某些软件工具的帮助下用来鉴定肽序列，这是 PSD 谱图一个很有价值的应用。

6.7 MALDI 的优缺点

虽然对于蛋白质组学分析来说没有十全十美的 MS 仪器，但是 MALDI-TOF MS 在以下四个方面是值得称道的。

第一，易于操作。仪器的主要部分是易于使用和稳定的。MALDI-TOF 仪器是最容易操作的 MS 仪器，这主要是因为没有难操作的 HPLC-MS 界面。这些仪器一般很适合于“随到随用”或“开放使用”。仪器系统不需预约，可供若干使用者日常使用。这样，在一个公用蛋白质组学设施中的一台 MALDI-TOF 仪器每天可很容易进行几百个分析。

第二，目前最广泛使用的 MALDI-TOF 仪器适合于自动化样品制备装置，这种新型自动化装置有利于高通量蛋白质的分析。2D 凝胶的制备和制图、蛋白质的获取和消化、将消化物放置到供分析的多样品 MALDI 靶子等工作都可通过自动化装置完成。这种结合不仅简化了高通量蛋白质组分析的工作，而且增加样品分析的速度和重复性。

第三，随着 TOF 分析器的精确性和分辨率的提高，所产生有用蛋白质组学数据的效率也得以提高。下一章中将看到，用 MALDI-TOF 数据通过肽质量谱进行可靠蛋白质鉴定的一个基本要求是对肽的精确质量测定。目前的 MALDI-TOF 仪器完全符合这项要求。

最后，MALDI-TOF MS 灵敏性高。MALDI-TOF 仪器能分析飞摩尔量的肽并提供高质量的 MS 数据，更好的仪器在最适条件下能有阿摩尔（10^{-18}）或更高的灵敏度。仪器操作的发展将提高其灵敏度、分辨率和质量精确性。

上文给出了关于 MALDI-TOF 仪器的各种优点，似乎没有什么理由考虑使用其他仪器了。然而，MALDI-TOF 也确实存在某些缺陷。

首先，这些仪器最适宜测定肽质量。这一类信息对蛋白质鉴定而言仍有其局限性。肽离子片段可提供有更大内在价值的真实序列数据，但是 MALDI-TOF 仪器不适于提供这类信息。如前所述，某些高端 MALDI-TOF 仪器可进行的 PSD 分析能够提供肽序列，但这不是真正的串联 MS 技术（见下文），不如 ESI 串联 MS 所提供的肽序列信息可靠。

第二，MALDI-TOF 分析能否成功在很大程度上依赖样品的质量。污染了

大量变性剂、缓冲液盐、金属或有机修饰物（如 DTT、尿素、甘油）的肽消化样品极大抑制在 MALDI 离子源的肽电离。由于没有在线 HPLC 系统从样品中去除污染物，MALDI 对各种污染相当敏感，污染因子可以影响所有 MS 分析。幸运的是研究者们使用固相清理工具已能够从分析样品中清除盐和其他污染物。

6.8 ESI 串联 MS 仪器

ESI 串联 MS（或 ESI-MS-MS）是电喷雾电离串联质谱的标准缩写。ESI 是指在仪器离子源中产生离子的过程。串联质谱是指可以进行多级质量分析的仪器。在 ESI-MS-MS 仪器中使用几种不同类型的质量分析器，最常用的是四极杆、离子阱和 TOF 质量分析器。在某些情况下，以不同的组合使用这些分析器。在研究分析蛋白质组学问题中，ESI 离子源与各种串联质量分析器的多重组合提高了仪器的灵活多用性。

6.9 溶液中的肽离子

为了理解 ESI 怎样工作，我们先讨论肽段样品溶液的酸-碱化学。在 MALDI 中，样品是肽和基质的结晶混合物。与之相反，用 ESI 分析的样品为水溶液中的肽或蛋白质。肽在溶液中以离子存在是因为肽片段上的某些功能基团在不同 pH 溶液中发生电离。羧酸在低于 pH3.0 时质子化（非电离的），溶液 pH 高于 5 时电离。相反，N 端胺和组氨酸氮是弱碱，溶液酸碱性在低于 pH7.0 时电离。赖氨酸和精氨酸的氮功能基团通常在低于 pH8.5 时电离。这意味着在酸性 pH（pH3.5 或更低）溶液中，胺的质子化使肽和蛋白质带有总的净正电荷。在碱性 pH 溶液中，胺和羧基的去质子化产生总的净负电荷。肽离子上的正电荷容易使肽裂解，而且在酸性 pH 下可以提高肽的 HPLC 层析特性。因此肽的 ESI 大都是在酸性样品的正离子模式中进行。

6.10 ESI 中肽离子电荷状态

ESI 的一个突出特征是使蛋白质和肽产生多电荷离子。许多肽具有多质子接受位点，在溶液中以单电荷或多电荷离子存在，尤其是由胰蛋白酶消化产生的肽片段，因为这些肽在 C 端有赖氨酸或精氨酸残基，并有 N 端氨基，它们在酸性溶液中质子化。蛋白质和肽的“多电荷”使形成的离子处在四极杆和离子阱分析器质量范围内，这些分析器的质量范围比 TOF 分析器的质量范围更小。例如，一个单质子化的 20 kDa 蛋白质（m/z = 20 001）的绝对质量超出四极杆质量分析器的质量范围，四极杆质量分析器的质量范围为 2 kDa，有时 4 kDa。一般的 20 kDa 蛋白质在溶液中可接受 10～30 个质子。溶液中的蛋白质分子有些接受 20 个质子，其 m/z 为 20 020/20 = 1 001；某些可接受 21 个质子，其 m/z 为 20 021/21 = 953；某些可接受 19 个质子，其 m/z 为 20 019/19 = 1 053 等等。很明显完整蛋白质的 ESI 质谱是所谓的“多电荷层”，代表溶液中蛋白质的所有

不同状态（图 6.4A）。电荷简化算法和软件可以将这种谱图转换成代表实际蛋白质质量的谱图（图 6.4B）。蛋白质存在多种电荷状态是因为这些大分子有多个可接受质子的受体位点，每一种受体与溶液平衡。

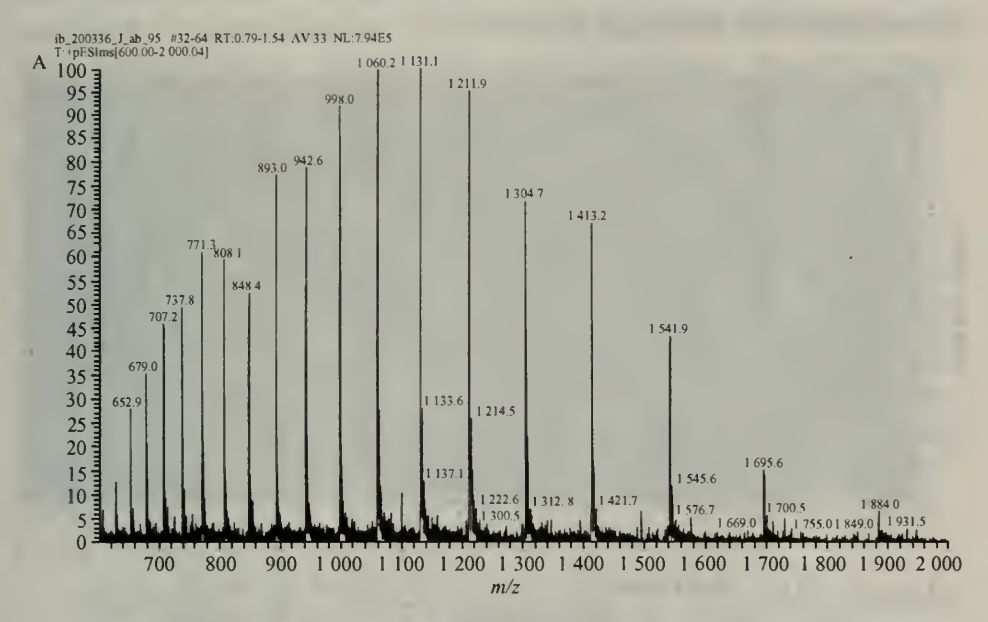


图 6.4A　牛脱辅基肌红蛋白的 ESI-MS 分析

来自蛋白质不同电荷形式信号的“多电荷层”。

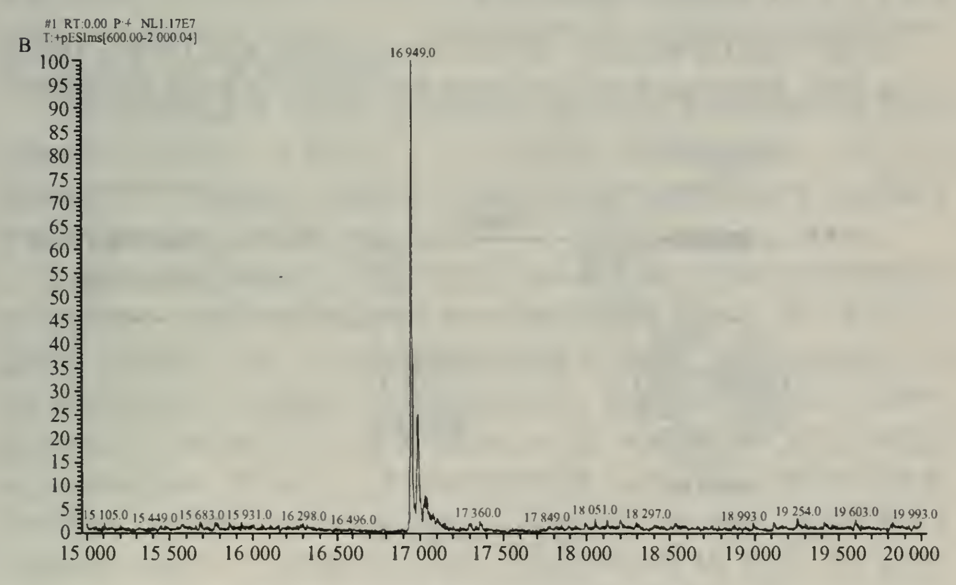


图 6.4B　牛脱辅基肌红蛋白的 ESI-MS 分析

去除环绕的谱图只有单个信号。

与完整蛋白质不同，250～2 500 Da 的肽以单、双和三电荷离子的混合物存在，这取决于肽的大小和存在的碱性氨基酸的数量。在这个质量范围的肽主要是双电荷离子，但也常可观察到单电荷和三电荷离子。在图 6.5 模型肽 AVAGCAGAR 的谱图中可看到这些离子的分布①。

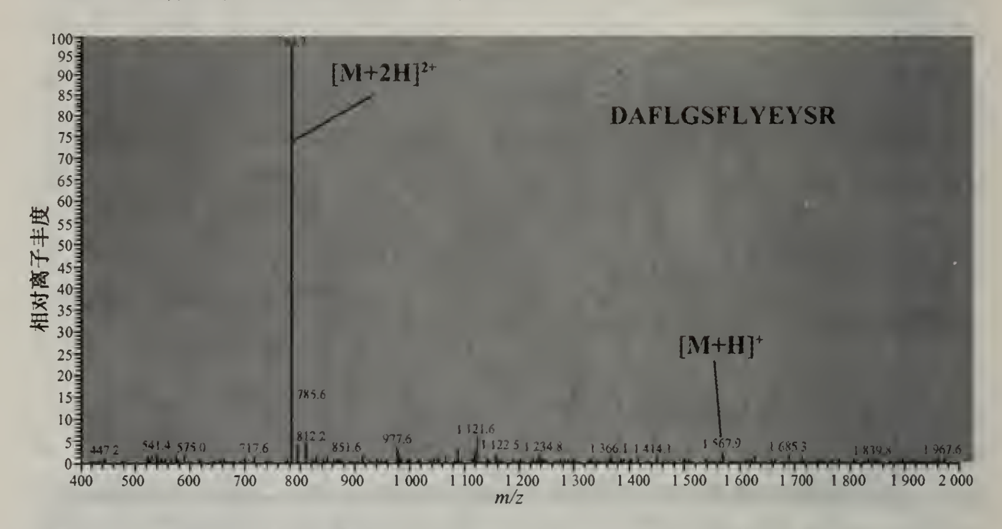


图 6.5　胰蛋白酶肽 DAFLGSFLYEYSR 的全扫描 ESI-MS 分析

展示一个 m/z 为 1567.9 的单电荷离子和一个 m/z 为 784.7 的双电荷离子。

6.11　ESI 离子源怎样工作

ESI 离子源的力学相对简单（图 6.6）。样品通过液流流动（通常从 HPLC）进入离子源，穿过有持续高电压的不锈钢锥体或进样针，随着液流离开进样针，

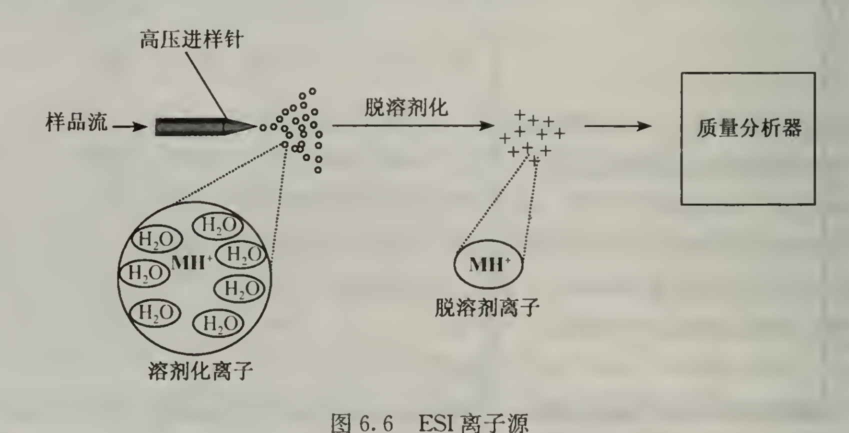


图 6.6　ESI 离子源

① 原文表述有误，图 6.5 实际上是肽 DAFLGSFLYEYSR 的谱图。——译者注

样品分散为细雾状微滴。微滴含有肽离子以及 HPLC 流动相组分（水、乙腈、醋酸等）。离子源需进一步从溶剂组分中分离肽离子，将离子转移到质量分析器。有两种方式可以从溶剂中分离肽离子。其一，在某些离子源中，微滴穿过一个加热的毛细管，完成脱溶剂化过程。其二，可以通过氮气幕穿过雾状微滴引起脱溶剂化。在这两种情况下，肽离子穿过离子源进入质量分析器，微滴中的大量溶剂通过真空系统泵出，肽离子被吸入质量分析器。

6.12 串联质量分析器

有三类串联质量分析器与 ESI 离子源联用进行蛋白质组分析。这些质量分析器是三级四极杆（常称为“triple quad”）、离子阱和四极杆飞行时间（Q-TOF）。这些质量分析器在某些工作细节上有所不同，但其基本原理相同。串联 MS 分析器从 ESI 离子源产生的肽离子混合物选择 m/z 相同的肽离子，然后进行碰撞诱导解离（CID）。CID 诱导使肽成为离子片段和中性片段。根据 m/z 分析离子片段，产生产物离子谱图。从串联或 MS-MS 谱图中含有的信息可以推导肽序列，也可以确定肽修饰的性质和修饰位点。

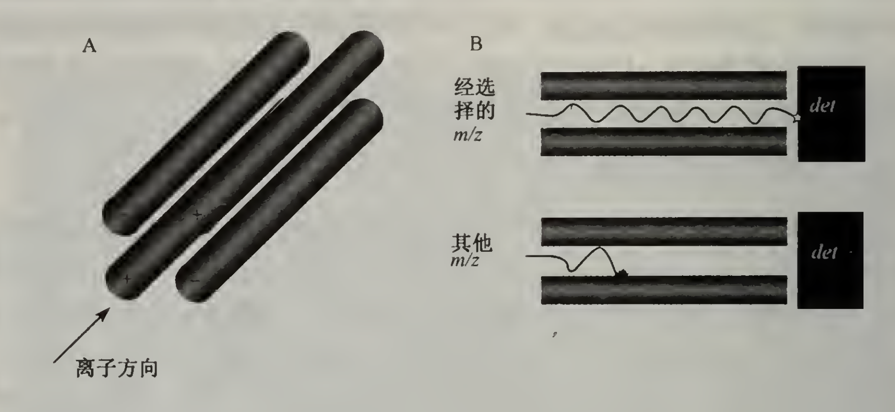
6.13 三级四极杆质量分析器

四极杆质量分析器由 4 个平行的金属圆棒组成（图 6.7A）。加到金属圆棒上的直流和射频电压形成磁场，使肽离子在圆棒之间沿轴方向以螺旋形轨道前进。圆棒上施加一定的电压允许某一特定 m/z 值的肽离子通过四极杆，而大于或小于这一 m/z 值的肽离子向外飞行，不能通过四极杆。通过迅速改变圆棒上的电压，可以分析 m/z 值不同的离子。

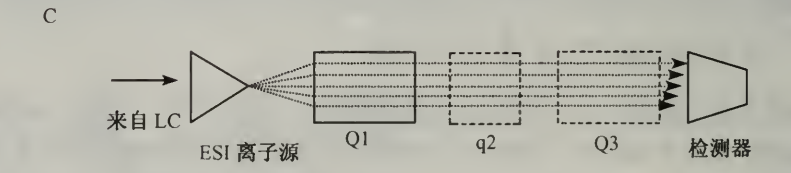
三级四极杆具有两套四极杆（Q1 和 Q3，图 6.7B），这两套四极杆由一个只受射频电压控制的另一个四极杆（q2，Q 小写是广泛接受的惯例）分开。中间的四极杆 q2 作为碰撞池。在碰撞池中肽离子和中性气体原子的碰撞导致肽离子裂解成为肽片段。检测器置于 Q3 之后。

三级四极杆以两种基本方式进行工作。第一种方式，Q1 快速扫描分析来自离子源的离子，记录在给定时间内来自离子源的全部离子的 m/z 值（图 6.7C）。这称为“全扫描”分析，产生离子源的全部肽离子（如单、双或三电荷离子）信号，这相当于在一定扫描时间（一般时间为 1 秒）进入离子源的肽离子“快照”。另一种工作方式使用 Q1 作为质量过滤器。在 Q1 中固定电压，只允许有特定 m/z值的离子通过（图 6.7D）。然后这些肽离子进入 q2，在 q2 与氩气原子碰撞，进行裂解。Q3 根据产生的片段离子的 m/z 进行分析。Q3 对规定质量范围反复扫描检测片段离子。三级四极杆用这种工作模式获取串联 MS 数据。用三级四极杆进行 MS-MS 分析的效率取决于被分析肽离子的性质和仪器的设定，包括在 q2 的 Ar 气压和用于 CID 的能量设定。在大多数三级四极杆的 MS-MS 实验中，实

际上进入 q2 的前体离子只有部分进行裂解，所发生的裂解范围有时要比在离子阱的裂解更广泛（见下文）。这样，三级四极杆的最佳 MS-MS 操作需要仔细调试仪器参数，以便得到最适程度的肽裂解。



C



D

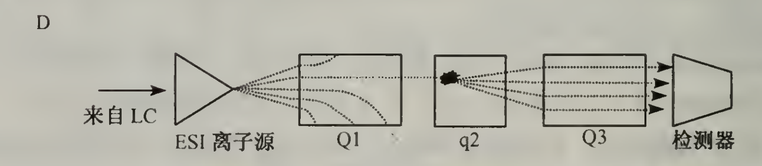


图 6.7　三级四极杆 MS 仪器的图示

A. 四极杆质量分析器；B. 选定的 m/z 值肽离子与其他 m/z 值肽离子的轨道；C. 三级四极杆在全扫描模式的操作；D. 在 MS-MS 模式中三级四极杆的操作。

在蛋白质组学的研究中，三级四极杆是用于串联 MS 的最早仪器。四极杆质量分析器允许选择特定肽离子（用 Q1），通过 MS-MS（用 Q3）进行离子片段的分析，其精确度达到真实 m/z 值的±0.5 amu，这种质量精确程度使得三级四极杆的 MS-MS 数据可直接用来解释氨基酸序列，也可以用 MS-MS 谱图与数据库蛋白质序列相关联的算法得到肽序列（见下文）。

6.14　离子阱质量分析器

离子阱质量分析器的设计和操作与三级四极杆有很大不同。三级四极杆是根

据肽离子穿过分析器的“飞行”来分析，而离子阱是收集和储存肽离子以便进行MS-MS分析。这种分析器设计简单。来自离子源的肽离子直接进入离子阱。离子阱由一个顶部电极、一个底部电极（端盖）和中间部位的一个环形电极组成（图6.8A）。离子阱本身约为一个葡萄柚大小。离子阱中收集到的肽离子通过DC（直流）和射频电压保存在离子阱中的小窝中。少量的氦用作“冷却气体”帮助控制肽离子的能量分布。在全扫描模式中，调节电极上射频电压，根据 m/z 值肽离子从离子阱中被连续逐出（图6.8B），这产生了在任何给定时间离子阱中所有肽离子的谱图。检测来自离子源的肽离子时，离子阱连续重复下列循环：①用肽离子填充离子阱；②根据 m/z 值通过扫描将离子逐出。与三级四极杆不一样，离子阱产生一系列间断的分析，而不是连续的分析。与三级四极杆类似，只要肽离子的 m/z 值在分析器的质量限定范围，离子阱可以检测ESI形成的多电荷肽离子。

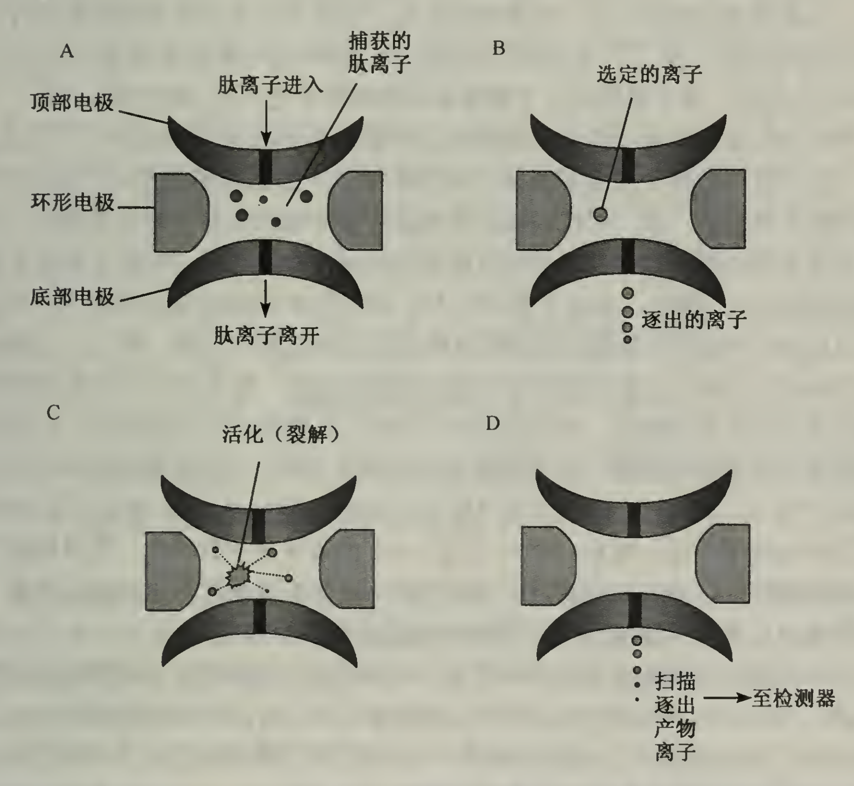


图6.8　离子阱MS仪器的图示

A. 在分析器中捕获离子；B. 不同 m/z 值肽离子的连续“扫描逐出”；C. 离子的碰撞诱导解离（裂解）；D.（C）中前体离子裂解产生的产物离子的连续“扫描逐出”。

进行MS-MS分析时，来自离子源的肽离子填充离子阱。然后选择某一感兴趣的特定肽离子，通过调节离子阱电压逐出所有其他 m/z 值的肽离子（图

6.8B)。再迅速增加离子阱电压以便增加存留肽离子的能量，导致肽离子与离子阱氦气原子的能量碰撞，诱导离子的裂解（图 6.8C)。然后捕获离子阱中的片段，根据不同的 m/z 值将其扫描逐出（图 6.8D)。

通过离子阱进行 MS 分析可以恰当比拟为"罐头盒中的石块"。可以用这个比拟总结离子阱 MS-MS 实验。将一小堆不同大小的石块放入一个罐头盒。选定其中的一块，把其余的石块扔掉。然后迅速用力摇动罐头盒，使选定的石块片段变成小石子。再依次取出石子进行称重。

离子阱的一个显著特征是 MS-MS 分析的肽片段离子可以保留在离子阱中，再进行一次裂解。第二次 MS-MS 分析的肽片段也可以保留，并进一步形成更小的肽片段。这一类分析称为 MS^n，在某些情况下可以产生非常详细的裂解信息。然而，由于以下的两个原因，阻碍了 MS^n 分析在蛋白质组学中的应用。第一，进行 MS-MS 分析时，由于不能确定在一个肽离子的 MS-MS 分析中会形成什么离子，也就很难选择用于进一步裂解的片段，因此目前还不能预期需要进行什么 MS-MS-MS 分析。第二，总肽离子数随 MS 循环的次数增多而减少。在一个 MS-MS 分析后，离子阱中往往不能有足够的肽离子进行有用的分析。

串联 MS 分析的离子阱和三级四极杆分析的不同之处还表现于以下几方面：第一，离子阱中肽离子裂解类型和三级四极杆产生的肽裂解类型有一定的差异。在通用操作条件下，离子阱比四极杆更倾向于诱导前体肽离子的完全裂解。这意味着在离子阱中更多前体肽离子能有效地转换成产物肽离子（因此可更有效地转换成序列信息)。确实，在离子阱 MS-MS 谱图中通常看不到前体肽离子信号，而三级四极杆 MS-MS 谱图中却能明显观看到前体肽离子信号。第二，三级四极杆在 MS-MS 分析中比离子阱能诱导更大范围的裂解。离子阱产生的大多数肽裂解都可直接用于序列测定，而三级四极杆 MS-MS 谱图有另外的特征。这些特征可解析意义不明确的数据，并可提供额外的细节。第三，离子阱 MS-MS 分析中的所谓"低 m/z 限制"。离子阱的 MS-MS 分析不能记录 m/z 值低于前体离子 25%的产物离子质量，即在一个 m/z 为 1 000 的前体离子分析中，可以检测的最低片段离子是 m/z 为 250 的离子。对肽 MS-MS 分析这通常不是问题，因为一般可以从相应较大肽片段的 m/z 值推出低质量肽片段的特征。

另外，离子阱质量分析极高的质量分辨率随离子扫描逐出和检测速度的增加而下降。在用于全扫描和 MS-MS 分析扫描速度下，离子阱能准确地解析 m/z 值相差至少 1 amu 的离子。在自动操作中，可使用对限定质量范围的慢全扫描，在 MS-MS 分析前精确测定离子的电荷状态。下面将看到，有关前体电荷状态的信息对用 MS-MS 数据测定肽序列是非常有帮助的。

6.15 自动数据获取

某些仪器控制软件允许三级四极杆或离子阱在全扫描和串联 MS 模式之间自

动切换获取肽 MS-MS 谱图。在这种方法中，仪器默认为全扫描模式来检测从离子源出现的肽离子。在检测时，仪器选择信号最强的肽离子进行 CID 以得到 MS-MS 谱图，然后仪器切回到全扫描模式，选择下一个信号最强的肽离子并进行 CID。这种切换反复循环，自动得到多个肽离子的 MS-MS 谱图（图 6.9）。这种自动化仪器控制方法称为数据依赖扫描或数据依赖 MS-MS，它非常适合在复杂肽混合物的 LC-MS-MS 分析中获取大量肽的 MS-MS 谱图。

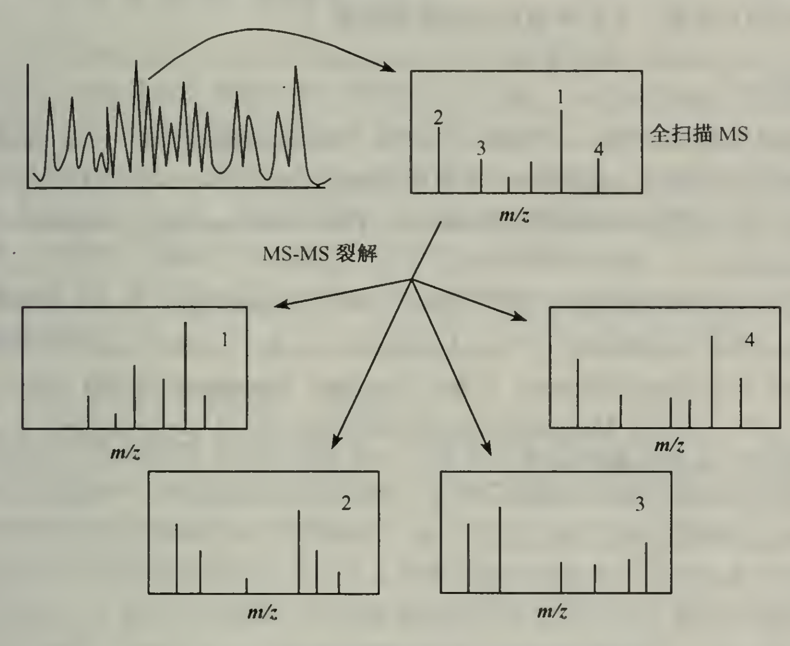


图 6.9　数据依赖扫描自动收集 MS-MS 谱图

6.16　其他质量分析器：Q-TOF 和傅里叶变换离子回旋加速共振 MS 仪器

有两种新质量分析器影响着分析蛋白质组学的发展，这两种仪器都使用 ESI 离子源。第一个是四极杆-飞行时间质量分析器，通常称为 Q-TOF，这是其两个组分的共同缩写。（应该指出混合缩写“Q-TOF”是一种商业仪器的商标名称。我在这使用缩写只是因为它的简洁和清晰，而不是对产品的认可。）Q-TOF 仪器在结构上 TOF 质量分析器代替了四极杆 Q3，在功能上与三级四极杆相同。最近 TOF 技术（前面已讨论）的改进使这些分析器的速度和分辨率都有很大提高。在 Q-TOF 中，除了 MS-MS 中的产物离子由 TOF 质量分析器而不是由三级四极杆 Q3 分析外，全扫描和 MS-MS 实验其他部分与三级四极杆工作方式相同。Q-TOF 的一个显著优点是 TOF 的质量分辨率远远高于四极杆，可以进行产物离子精确的质量测定，有利于从 MS-MS 谱图得到精确的序列测定。此外，TOF 的

较高分辨率和质量精确性产生的数据可更有效地用于软件辅助的数据解释。

傅里叶变换离子回旋加速共振 MS（称为 FT-ICR 或最常称为 FT-MS）某种程度上与离子阱类似。质量分析器使用强大磁场（典型磁场为 3～7 Tesla）和傅里叶变换算法同时检测所有在离子阱的离子。这些仪器可以用 ESI 离子源操作，即使分析非常复杂的肽混合物也能得到极高的分辨率。对分析蛋白质组学 FT-MS 是潜在的功能强大的工具。然而，FT-MS 是非常昂贵、不太容易操作的仪器，这些因素限制了它们对蛋白质组学的影响。

推荐读物

Jonscher，K. R. and Yates，J. R. （1997） The quadrupole ion trap mass spectrometer：a small solution to a big challenge. *Anal. Biochem*. **244**，1—15.

Siuzdak，G. （1996） *Mass Spectrometry for Biotechnology*. Academic Press，San Diego.

Stahl，D. C.，Swiderek，K. M.，Davis，M. T.，and Lee，T. D. （1995） Data-controlled automation of liquid chromatography/tandem mass spectrometry analysis of peptide mixtures. *J. Am. Soc. Mass Spectrom*. **7**，532—540.

Yates，J. R. （1998） Mass spectrometry and the age of the proteome. *J. Mass. Spectrom*. **33**，1—19.

7 用肽质量指纹谱鉴定蛋白质

7.1 什么是肽质量指纹谱

肽质量指纹谱是用MS测定蛋白水解肽片段质量的蛋白质鉴定技术。用实验测得的蛋白质酶解肽段质量在蛋白质数据库中检索，寻找具有相似质量的肽，从而鉴定蛋白质。肽质量指纹谱很适合分析蛋白质组学，因为它将简单的方法与稳定的高通量仪器操作（特别是MALDI-TOF MS）相结合。与其他基于MS的分析蛋白质组学技术一样，用肽质量指纹谱进行蛋白质鉴定的质量取决于MS数据的质量、数据库的准确性以及使用的检索算法和软件的功能。在本章将详细讨论肽质量指纹谱。我们将描述怎样用测定的肽质量鉴定蛋白质以及使鉴定过程自动化的算法和软件。

7.2 肽质量指纹谱：概述

获得某个生物的整个蛋白质组后，用胰蛋白酶切割成不同大小的胰蛋白酶肽。胰蛋白酶专一性较高地切割蛋白质，它只在赖氨酸和精氨酸残基切割（当精氨酸后的氨基酸是脯氨酸时不能切割）。每一个蛋白质用胰蛋白酶消化产生一定数量的有特定长度和序列的肽，而且这些肽有特定的质量。只要胰蛋白酶肽段混合物中的每一个肽与其来源蛋白质和氨基酸序列位点相关联，蛋白质组的全部信息就得以保留。当然，并不一定要在实验中进行胰蛋白酶消化。可以通过计算机在数据库中虚拟消化全部蛋白质，产生肽的总目录，通过将核苷酸序列信息转换成蛋白质序列信息然后进行消化也可以达到同一目的。理论上，测序正确的完整基因组序列可以产生完整的蛋白质目录，所以也能产生完整的胰蛋白酶肽的目录。

肽段的这种超级目录已成为很有价值的参考工具。目录按质量大小从低到高排列这些胰蛋白酶肽，核查这种目录会揭示长度超过6氨基酸的某些肽（约700 Da）在目录中有惟一质量。

鉴定某个生物的未知蛋白质，首先用胰蛋白酶消化蛋白质产生胰蛋白酶肽。消化后得到的每一肽段的质量各不相同。如果每一个胰蛋白酶肽消化产物的确切质量已知，那么取一定质量的胰蛋白酶肽段与整个目录进行比较，我们会发现目录中有与之质量完全一样的肽片段（图7.1）。如果在全部肽质量的目录中 这种匹配是惟一的，我们几乎可以肯定这两个肽段是相同的。在目录中与之匹配的肽段的序列位置和来源已知，从而可以肯定我们的胰蛋白酶肽与之来自相同的蛋白质。然后可以用未知蛋白质的第二个胰蛋白酶肽段以相同方式与目录匹配，从目

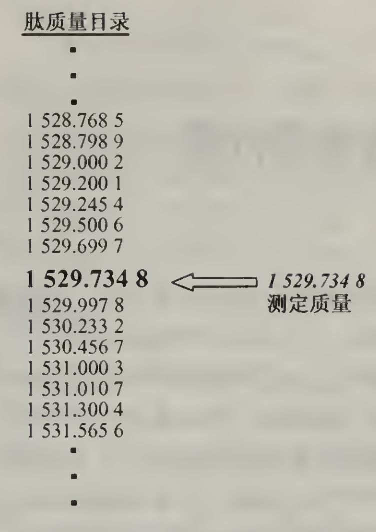


图 7.1　肽 m/z 值与蛋白质序列数据库中肽离子质量目录的匹配

录中得到与未知肽段匹配的已知肽片段和蛋白质。未知蛋白质的胰蛋白酶肽段质量与目录中胰蛋白酶肽段的多个匹配可说明未知蛋白质的特性。即使肽质量目录中有多个条目与未知胰蛋白酶肽段匹配，可以根据未知蛋白质的多个肽段与目录中已知的同一个蛋白质的多个肽段相匹配来确定未知蛋白质。只要能使肽质量与一个好的质量目录匹配，就能简单地通过测定胰蛋白酶肽质量鉴定未知蛋白质。这是用肽质量指纹谱鉴定蛋白质的基本要点。

当然，这是一个高度理想化的例子。以下两点决定了肽质量指纹谱能否成功地用于鉴定蛋白质。第一，必须能够精确测定肽质量。第二，必须有有效且精确的蛋白质序列数据库。

7.3　肽质量指纹谱：分析方法

肽质量指纹谱方法相对简单。使用以可预测方式切割蛋白质的特定蛋白酶（最常用的是胰蛋白酶）处理蛋白质样品。除胰蛋白酶外，其他一些蛋白酶，甚至化学试剂也可特异性地将蛋白质切割成肽。不管用什么酶消化蛋白质关键是产生专一性切割，因为蛋白质序列数据库中也进行相同的切割，产生用于匹配的肽质量目录。然后用 MS 分析肽，进行质量测定。通常仪器实际上测定的是肽段 m/z 值，它可以被转换成质量。理论上，任何 MS 仪器都可用来测定肽的 m/z 值，但是用肽质量指纹谱进行可靠的蛋白质鉴定时需要高度精确的质量测定。下面举一个实际例子来说明这一点的重要性。

人血红蛋白 α 链的胰蛋白酶消化产生 14 个胰蛋白酶肽，其中序列为 VGAHAGEYGAEALER 肽的单个同位素精确质量为 1 528.734 8 Da，其单电荷离子的 m/z 值为 1 529.734 8，这个肽对 SWISS-PROT 数据库中所有小鼠和人蛋白质的检索结果如表 7.1 所示。

表 7.1　质量精确性和质量误差极限对肽质量指纹谱检索结果的影响[a]

检索 m/z	质量误差极限/Da	命中数
1 529	1	478
1 529.7	0.1	164
1 529.73	0.01	25
1 529.734	0.001	4
1 529.734 8	0.000 1	2

a. 在 http://prospector.ucsf.edu 网址用 MS-FIT 程序进行检索。

如果质量误差极限为 1 Da（即测定质量在真实值的±1 Da 范围内），对最接近的整数 m/z 值（即测定的 m/z 值为 1 529）进行测定，有 478 个匹配。换句话说，有 478 个来自小鼠和人蛋白质的胰蛋白酶肽为 1 529±1 Da。如果能对肽离子进行更精确的 m/z 值测定和使用更严格的质量误差极限，可以把最终匹配的数目缩小到两个肽。有趣的是，这两个肽是来自人血红蛋白 α 的 VGAHAGEYGAEALER 和来自小鼠血红蛋白 α 的 IGGHGAEYGAEALER。虽然这两个肽有 4 个氨基酸的不同，但是它们的 m/z 值都是 1 529. 734 8。

这里要强调的是更精确的 m/z 值测定为肽质量指纹谱提供更可靠的数据。肽质量指纹谱的质量测定所用的 MS 仪器操作必须能测定在实际值±0. 05 Da 范围内的肽质量。具有延迟提取和反射器分析器的现代 MALDI-TOF 仪器可做到这一点，并已被广泛应用到这一领域。然而用单一 m/z 值 1 529. 73 和±0. 01 Da 的质量误差极限仍产生 25 个与之匹配的蛋白质（表 7. 1）。与之匹配的某些蛋白质列于表 7. 2。

表 7. 2　与 m/z 值为 1 529. 73 的肽的质量指纹谱匹配的蛋白质

肽序列	鉴定	检索质量中匹配的 m/z（差异）
IGGHGAEYGAEALER	小鼠 Hb α	1 529. 734 8（−0. 004 8）
VGAHAGEYGAEALER	人 Hb α	1 529. 734 8（−0. 004 8）
MGTGWEGMYRTLK	小鼠晶状体上皮细胞蛋白质 LEP503	1 529. 724 5（0. 005 5）
MADEEKLPPGWEK	人 PIN1 类蛋白	1 529. 731 0（−0. 001 0）
DTQTSITDSSAIYK	小鼠信号识别颗粒受体 β 亚单位	1 529. 733 5（−0. 003 5）
NDSSPNPVYQPPSK	小鼠过氧化物酶体组装因子 1	1 529. 723 6（0. 006 4）
MNLSLNDAYDFVK	人双专一性蛋白磷酸酶	1 529. 731 0（0. 001 0）

所有这些匹配的蛋白质完全是在 0. 01 Da 的特定误差极限之内。在这个标准上，这些匹配蛋白质中的任何一个都可能是我们所要研究的蛋白质。怎样从这些极为相似的匹配中鉴定出正确的蛋白质?

答案是准确的蛋白质鉴定通常需要多个肽匹配。在表 7. 1 的例子中，即使最精确的质量匹配也不能确定我们所分析的肽是来自人血红蛋白或来自小鼠血红蛋白。蛋白质样品的胰蛋白酶消化能产生多个肽段，由此得到多个可供数据库检索的 m/z 值。表 7. 3 表明增加进行检索的肽 m/z 值数目的优点。

表 7. 3　多个肽质量对用肽质量指纹谱进行蛋白质鉴定的影响[a]

检索 m/z[b]	质量误差极限	命中数
1 529. 73	0. 1	204
1 529. 73		
1 252. 70	0. 1	7
1 529. 73		
1 252. 70		
1 833. 88	0. 1	1

a. 在 http：//prospector. ucsf. edu 网址用 MS-FIT 程序进行检索。

b. 实际肽 m/z 值是 1 529. 734 8（VGAHAGEYGAEALER），1 252. 707 4（FLASVSTVLTSK）和 1 833. 884 5（TYFPHFDLSHGSAQVK）。

用人血红蛋白 α 链的一个或两个肽段进行检索产生多个与之匹配的已知肽段。用人血红蛋白 α 链的 14 个胰蛋白酶肽中的 3 个肽的 m/z 值进行检索产生一个与之对应的匹配。

7.4 肽质量指纹谱：复杂情况

前面用一个非常简单的例子解释用肽质量指纹谱鉴定蛋白质的概念。用两个或三个肽的 m/z 测定来鉴定如人血红蛋白 α 链那样的蛋白质前体相对简单。当然，在这个例子中我们假定肽质量测定是正确的且误差达到 0.01 Da 或更精确。这些例子证明更精确的和对多个肽的质量测定极大提高了鉴定的准确性。根据仅仅来自两个或三个肽的数据可以通过“手工”成功构建肽质量指纹谱。

在实际工作中有以下几个因素使肽质量指纹谱变得较为复杂。首先，肽的真实数据不像前面的例子那样完美。虽然装备了反射器或延迟提取的 MALDI-TOF 仪器能够在 0.005 单位或更精确的范围内测定肽离子的 m/z 值，但是误差仍是不可避免。第二，在实际样品的 MALDI-TOF 谱图中经常有多个信号，其中许多信号是由于其他蛋白质的混入。必须考虑 2D 凝胶的大多数样品点含有 2～3 个蛋白质，一个典型的 50 kDa 蛋白质产生 25～40 个胰蛋白酶肽，其他样品污染（如不小心操作污染的人角蛋白）也会在谱图中产生信号。这些因素导致产生复杂的、代表多个蛋白质肽片段的谱图。第三，由于偶然因素而不是实际特征导致某些数据库匹配的可能性总是存在的。较大蛋白质的假阳性匹配可能性更大，这主要是因为它们比较小蛋白质产生更多胰蛋白酶肽。

7.5 肽质量指纹谱的软件工具：寻找匹配

面对大量的相关数据和计算似乎令人不知所措，但我们可以寻求数据缩减算法和软件的帮助。有若干软件工具用于肽质量指纹谱鉴定蛋白质，其中有些软件列在这一章后面。下面简单介绍这些程序可做哪些工作。

程序使用者首先考虑应选择什么样的检索数据库。可以选择蛋白质或基因序列数据库（如果选择后者，要翻译基因序列）。SWISS-PROT 蛋白质序列数据库已被广泛应用。其他常用蛋白质序列数据库有 OWL 和 NCBInr 数据库。程序使用者提供有关样品来源的信息，限定对相关生物的检索。例如，来自小鼠蛋白质的样品可以对所有生物、对哺乳类序列、对啮齿目或对小鼠进行专一的序列检索，这样可以减少与数据进行比对的数目，降低在其他生物中假阳性匹配的数目。除了这些特点，程序使用者也可提供要检索蛋白质的分子质量范围，这也可降低需进行比对的数目。

第二步，使用者表明用来切割蛋白质的酶，指定“漏掉的切割”的可能数目，这些漏掉的切割来自酶的不完全消化，匹配算法可以产生这类肽的名单，以防它们在样品中存在。第三步，程序使用者指定在匹配算法中要考虑的一些肽修饰。例如，胰蛋白酶消化方法通常包括用碘乙酰胺或碘乙酸进行的半胱氨酸巯基

的还原和烷化，这种操作会改变肽的半胱氨酸残基的质量。此外，自由半胱氨酸巯基可能在 SDS-PAGE 中被丙烯酰胺修饰。程序使用者也可指定一些常见的蛋白修饰，如磷酸化、硫化、糖基化和 N 端修饰。所定义的这些修饰允许程序在数据库中产生修饰前和修饰后肽质量的匹配，因而使特定肽段在修饰前后的 MS 数据与数据库中的条目匹配。第四步，程序使用者输入 MS 测定的 m/z 值或自动指定要估算的 MS 数据文件夹。然后输入要求的质量误差极限以便控制“命中”所需要的 MS m/z 值和计算的 m/z 值的接近程度。

一旦程序使用者点击“Go”，软件从预过滤要使用的数据库开始工作。例如，如果指定小鼠为检索物种，所有非小鼠条目都被排除；如果选择的蛋白质质量范围为 2 000～100 000，所有在这个质量范围之外的蛋白质都被排除。然后软件程序用指定的酶对数据库中保留的序列进行虚拟消化，如果允许漏掉切割，肽片段目录中包括酶不完全消化产生的肽段。软件程序根据所指定的修饰产生相对应的肽片段。接下来程序按质量（或 m/z 值）大小排列整个肽目录，用谱图中每个 m/z 信号与这个目录进行比对。在程序使用者指定的质量误差极限内的所有匹配都被记录为“命中”，并用来计算相关蛋白质的得分和进行鉴定。

7.6 肽质量指纹谱的软件工具：给结果评分

在实际样品的 MALDI-TOF 谱图中一般有几十个 m/z 信号。肽质量指纹谱软件通常将所有的 m/z 信号与数据库中的某些条目匹配。然而，由于 m/z 测定误差、常有的样品污染和意料不到的翻译后修饰，不是所有匹配都指向相同的蛋白质。怎样评价命中以便决定哪一个蛋白质与谱图中的数据匹配得最好？

最简单的方法是把最高分给这样的蛋白质：它的预测的胰蛋白酶肽与最大数量 MS 数据的 m/z 信号匹配。如果仅检索一个 m/z 值，几个蛋白质会匹配得同样好。然而，随着检索更大数量的 m/z 值，对一个特定的蛋白质会有更多的匹配，使那个蛋白质相对其他蛋白质得分更高。使用极好的 MS 数据，这个相当简单的方法很有效。不过它倾向于给较大蛋白质较高分。如前面提到的，较大蛋白质产生更多胰蛋白酶肽，因此较大蛋白质与这些胰蛋白酶肽之一匹配的机会要大于较小蛋白质。

为解决这些问题，几个肽质量指纹谱程序使用更为复杂的得分算法。这些算法可校正由于蛋白质大小不同引起的得分偏倚（较大的蛋白质产生较多的肽），也可校正数据库中较小的肽与用于检索的 m/z 值有较多匹配的倾向，另外某些算法使用概率统计以便更好地定义蛋白质鉴定的意义。在本书写作时，用于肽质量指纹谱的基本工具可分为三组。

- 第一代免费和付费软件工具，这些软件根据谱图中与数据库中 m/z 值（在给定质量误差极限之内）相匹配的数目给出得分。包括 PepSea（http://www.protana.com）和 PeptIdent/MultIdent（http://www.expasy.ch/tools/pep-

tident. html)。

• 第二代免费和付费软件工具，这些软件使用的得分算法考虑到蛋白质大小和肽片段长度对匹配几率的影响。包括 MOWSE(http://srs.hgmp.mrc.ac.uk/cgibn/mowse)和 MS-Fit(http://prospector.ucsf.edu)。

• 第三代软件更多地使用基于概率的得分，提供得分的统计基础，估计某些匹配可能反映随机事件而不是真实特性的概率。这些程序包括 ProFound (http://prowl.rockefeller.edu/cgi-bin/ProFound) 和 Mascot(http://www.matrixscience.com)。

7.7 肽质量指纹谱：评价和展望

对于蛋白质鉴定，肽质量指纹谱方法有很多地方值得推荐，它在蛋白质组学中最接近"高通量"。在样品制备、MS 分析和数据缩减自动化的帮助下，用单一系统一天可鉴定几百个蛋白质。仪器操作（典型的如 MALDI-TOF）简易、稳定且灵敏。蛋白质和核苷酸序列数据库的迅速发展为数据库检索算法提供了更可靠的平台。检索算法的改进和复杂统计学方法的应用提高了蛋白质鉴定的可靠性。

然而，肽质量指纹谱也有一些局限性。首先，人类和其他研究较为深入和广泛的物种基因组和蛋白质序列数据库仍不十分完善和准确，这降低了可获得的匹配数据的质量，即使极好的 MS 数据和软件也对此无能为力。这种状况肯定会改进，但在近期仍是主要限制因素。第二，高等生物中大量高度同源蛋白质使得这些蛋白质的区分变得复杂。这些同源蛋白质的肽图谱非常复杂。肽质量指纹谱在酵母中可能较容易，但在小鼠和人中要复杂得多。第三，肽质量指纹谱主要是一种蛋白质鉴定技术。我们后面会看到，在蛋白质组学应用中，肽序列和肽修饰位点的信息是必要的，而我们不能从肽质量测定推出这些信息。

尽管有这些局限性，肽质量指纹谱仍是任何正规蛋白质组学实验室的一项必备的研究技术。基于 MALDI-TOF 的肽质量指纹谱的许多局限性可用 ESI 串联 MS 克服。下面两章我们将讨论 ESI 串联 MS。

推荐读物

Fenyo, D. (2000) Identifying the proteome: software tools. *Curr. Opin. Biotechnol.* **11**, 391—395.

Gras, R., Muller, M., Gasteiger, E., Gay, S., Binz, P. A., Bienvenut, W., et al. (1999) Improving protein identification from peptide mass fingerprinting through a parameterized multi-level scoring algorithm and an optimized peak detection. *Electrophoresis* **20**, 3535—3550.

Jensen, O. N., Podtelejnikov, A. V., and Mann, M. (1997) Identification of the components of simple protein mixtures by high-accuracy peptide mass mapping and database searching. *Anal. Chem.* **69**, 4741—4750.

8 用串联质谱分析肽序列

8.1 图型中有什么?

上一章我们讨论了用人血红蛋白 α 链的胰蛋白酶肽 VGAHAGEYGAEALER 的 m/z 值测定来鉴定蛋白质。在检索人和小鼠蛋白质序列数据库时，有两个肽与之非常匹配，它们的 $[M+H]^+$ 离子 m/z 值都是 1 529.734 8。这两个肽都与在小鼠和人之间高度保守的血红蛋白 α 肽相关。

人：VGAHAGEYGAEALER

小鼠：IGGHGAEYGAEALER

浏览这两个序列会发现"IGG"和"VGA"是不同的。进一步仔细的观察可发现"HGA"和"HAG"很接近，但不相同。我们看到的是一个图型，它能区分在某一方面（质量）一致，但序列明显不同的两个物种。这一章将讨论串联 MS 分析如何诱导肽裂解，在 MS-MS 谱图中肽裂解如何产生产物离子和怎样从 MS-MS 谱图的裂解图型测定肽序列。

8.2 肽序列中有什么?

使用字母表示法可以表明两个相同质量肽的图型和结构的不同。然而，MS 仪器测定肽及其 m/z 值，把结构还原成数字图型。理解肽结构的数字图型系统对理解串联 MS 谱图含有什么信息是必要的。我们以肽 AVAGCAGAR 为例，解释串联 MS 裂解。图 8.1 是这个肽的一级结构，以线性方式描述序列。肽是通

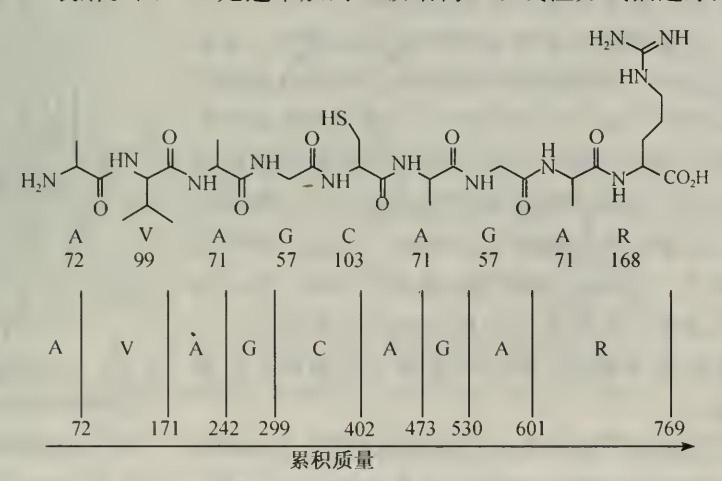


图 8.1 用质量不同的氨基酸构建的肽 AVAGCAGAR

表 8.1　氨基酸的平均残基质量

氨基酸	字母符号	平均残基质量
甘氨酸	G	57.05
丙氨酸	A	71.08
丝氨酸	S	87.08
脯氨酸	P	97.12
缬氨酸	V	99.13
苏氨酸	T	101.11
半胱氨酸	C	103.14
亮氨酸	L	113.16
异亮氨酸	I	113.16
天冬酰胺	N	114.10
天冬氨酸	D	115.09
赖氨酸	K	128.17
谷氨酰胺	Q	128.13
谷氨酸	E	129.12
甲硫氨酸	M	131.19
组氨酸	H	137.14
苯丙氨酸	F	147.18
精氨酸	R	156.19
酪氨酸	Y	163.18
色氨酸	W	186.21

过氨基酸末端失水缩合形成肽键的。图 8.1 中 AVAGCAGAR 的氨基酸残基由虚线表示。每一个残基在一端有一个酰胺 NH 基团，在另一端有一个 C=O 基团，带有一个质子的 α 碳在中间。决定每个氨基酸特定化学性质的侧链与 α 碳连接。含有上述基本组成的氨基酸单位称为残基，表 8.1 列出了常见氨基酸的鉴别和残基质量。

通过这种表格将氨基酸字母名称转换成氨基酸残基质量，我们可以将一系列氨基酸表示为一系列数字，这些数字与氨基酸残基的质量一致。另外必须在 N 端残基加上一个额外的质子（1 amu），在 C 端氨基酸加上一个额外的 OH（17 amu）。现在可以将 AVAGCAGAR 表示为一系列逐渐累加的数字序列，如图 8.1 下半部分所示。数字系列最确切代表了 MS 仪器如何处理这个肽和它的结构。

8.3　MS-MS 中的肽离子裂解

当肽离子在三级四极杆或 Q-TOF 的碰撞池中或在离子阱中与中性气体原子碰撞时，离子吸收动能诱导裂解。肽片段中的许多键都有可能断裂，但主要是肽骨架的切割（图 8.2）。一般使用被广泛接受的术语来描述肽离子的裂解。通常观察到的切割中，在羰基氧和酰胺氮之间键的裂解形成一个“y 离子”和一个“b 离子”。y 离子是电荷保留在源肽离子 C 端的片段，b 离子是电荷保留在源肽离子 N 端的片段。双电荷离子最有可能在分子的两端带电荷。当双电荷肽离子断裂成片段时，形成 b 离子和相应的 y 离子。当单电荷离子断裂成片段时，形成 b 离子或者 y 离子，肽的另一半作为中性片段丢失。很明显与单电荷离子相比，从双电荷离子的裂解可以得到双倍信息。

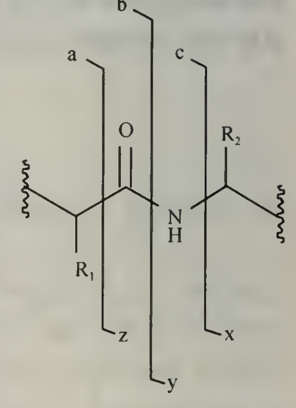


图 8.2　肽离子裂解产生的各种碎片离子的图示

图 8.2 也指出了肽骨架的其他裂解形式。在离子阱、三级四极杆和 Q-TOF 仪器得到的 MS-MS 谱图中偶尔也能观察到 a、c、x 和 z 离子，但它们不常出现，

因为这些裂解需要的能量多于产生 b 和 y 离子裂解时所需要的能量。（在磁分区仪器的串联 MS 分析中常可观察到这些离子。磁分区仪器在肽离子的碰撞诱导解离中使用更多的能量。）

8.4 MS-MS 谱图

要更好地理解 b 和 y 离子的裂解怎样产生特定图型，研究肽谱图是有帮助的。图 8.3 指出了预测的模型肽 AVAGCAGAR 的 b 和 y 离子裂解。图 8.4 是双电荷离子 AVAGCAGAR 的实际 MS-MS 谱图。从 N 端（左侧）逐渐向 C 端（右侧），可看到裂解产生了一系列 m/z 值逐渐升高的肽离子片段（b 系列）和一系列互补的 m/z 值逐渐下降的肽离子片段（y 系列）。图 8.3 还指出了由 G-C 键裂解产生的 b_4和 y_5离子片段。每个片段带有单电荷。图示片段中的质子化位点与源肽离子片段的质子化位点略有不同。一般认为这些结构很可能是由于碰撞诱导肽裂解而存在于气相的离子种类。然而，肽链不同位点质子化形成的多种离子形式也共同存在。在碰撞诱导解离中，双电荷前体中质子向肽酰胺氮的迁移可能帮助邻近肽键的切割。

图 8.3 肽 AVAGCAGAR 的可能的 b 和 y 离子片段

图为在甘氨酸和半胱氨酸之间切割产生的 b_4和 y_5离子结构。

在 MS-MS 谱图中从 b 和 y 离子系列可以确定出源肽的序列。先讨论 y 离子系列，在以上的例子中，在标记的 y_7和 y_6离子之间 m/z 值之差是 71 amu，这相

当于一个丙氨酸的残基质量。在 y_6 和 y_5 之间 m/z 值之差是 57 amu，这相当于一个甘氨酸。在 y_5 和 y_4 之间的差值是 103 amu，相当于半胱氨酸。仅用 y 离子系列的一小部分就可以建立在肽中存在一个“AGC”基序。从完整的 y 离子系列（y_8 到 y_1）可推测出“VAGCAGAR”基序。

b 离子系列与 y 离子系列互补，在 b_7 和 b_6 离子之间 m/z 的差值是 57 amu，相当于甘氨酸。在 b_6 和 b_5 之间的差值是 71 amu，相当于丙氨酸。完整的 b 离子系列（b_8 到 b_1）与“AVAGCAGA”基序相对应。y 和 b 离子系列从两个不同方向描述相同氨基酸序列，因而 AVAGCAGAR 双电荷离子的 MS-MS 谱图（图 8.4）中 b 和 y 离子的标定可提供准确的序列。

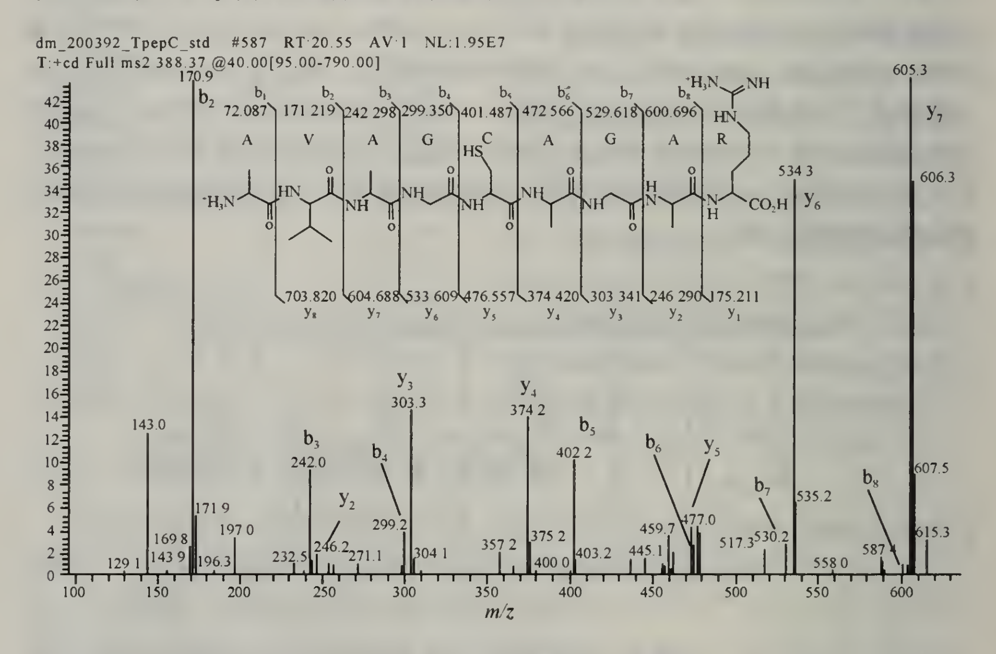


图 8.4　显示 b 和 y 离子的 AVAGCAGAR 的 $[M+2H]^{2+}$ 离子的加注释的 MS-MS 谱图

当然，我们所用的为已知肽序列的例子。在实际实验中，我们得到的只是肽序列未知的 MS-MS 谱图。然后我们需使用谱图、计算器和氨基酸残基质量表（表 8.1），鉴定谱图中的 b 和 y 离子以便推定肽序列，这称为从头测序。由于研究人员经验的不同和数据质量的好坏不同，一张谱图需要 15 分钟到几个小时来解释。单个 LC 串联 MS 分析往往产生成百上千的 MS-MS 谱图，所有这些谱图可能代表不同的肽序列，因此这种从头测序不是很好的方法，这种手工的从头测序无法应对所产生的海量数据。

为解决这一问题，已经开发出数据缩减算法和软件工具，能将 MS-MS 数据与数据库中肽序列进行比较，鉴定产生肽的蛋白质。这些程序包括 Sequest 和几种其他类似工具，将在后面详细描述。下面我们先对 MS-MS 谱图的其他特点以

及在肽 MS-MS 分析中常遇到的问题和异常情况进行讨论。

8.5 问题、独特性和脯氨酸

图 8.4 中 AVAGCAGAR 肽的 MS-MS 谱图与 LC 串联 MS 分析中常见的 MS-MS 谱图很接近。然而，不是所有实验中获得的谱图都有这么理想。受仪器灵敏度的限制，谱图可能是不完全的，很难解释的。即使需要分析的肽的质量完全在检测极限之内，仍然有一些影响因素使仪器不能产生具有完整的 b 和 y 离子系列的“完美”MS-MS 谱图。这些因素包括：①不同肽键产生片段的倾向性不同；②某些氨基酸独有的裂解特征；③脯氨酸对肽离子裂解的弱化影响。

现在仍不十分清楚影响不同肽键断裂难易的因素。但是我们已知道裂解确实依赖于质子化肽离子中质子迁移到不同肽链的酰胺氮的难易程度。容易质子化的位点最容易发生裂解。酸性氨基酸侧链对稳定正电荷也有一定作用，因而邻近谷氨酸或天冬氨酸残基的裂解常产生信号强的离子片段。在许多肽 MS-MS 谱图中，信号最强的离子片段常常是靠近源肽中间的裂解产生的。当肽离子在碰撞中获得能量时，能量在裂解过程中产生竞争性分散，易于裂解的位点会减少其他位点的裂解，使得某些片段离子减少或缺失。

胰蛋白酶肽离子的裂解图型特别重要，因为蛋白质组学研究主要采用胰蛋白酶消化蛋白质的 MS 分析，如前所述，胰蛋白酶肽常带有双电荷，因为它们在 C 端有赖氨酸或精氨酸残基。在胰蛋白酶肽的 MS-MS 谱图中，y 离子系列通常比 b 离子系列信号强，这是因为赖氨酸和精氨酸残基的碱性侧链有在肽片段 C 端保留正电荷的能力。值得注意的是某些双电荷肽离子的裂解产生一个双电荷产物离子和一个中性片段。这样，某些 MS-MS 谱图中的产物离子可能来自双电荷片段，而不是来自单电荷片段。这些情况不经常发生，然而它们有时干扰谱图的解释。

某些氨基酸侧链产生与肽骨架无关的特定裂解切割。例如，丝氨酸和苏氨酸残基很容易在羟基的侧链上脱水。脱水后产生的离子有时会比含有完整丝氨酸或苏氨酸的离子片段在谱图中的信号更强。磷酸丝氨酸和磷酸苏氨酸残基可发生类似的去磷酸（H_3PO_4）化。磷酸丢失后形成的离子常常可以在 MS-MS 谱图中占据多数，这经常是判断磷酸肽的可靠指标。其他的特定侧链丢失包括半胱氨酸 H_2S 的丢失以及谷氨酰胺和天冬酰胺残基的氮丢失。

在肽裂解中引起信号混淆的另一个因素是脯氨酸残基的存在。如在第 5 章提到的，当脯氨酸位于赖氨酸或精氨酸的 C 端一侧时，会阻碍胰蛋白酶切割。除了对消化的影响，脯氨酸也影响 MS-MS 分析中肽离子的裂解。脯氨酸残基肽键相对不易发生裂解，这是由于脯氨酸残基具有特有的结构，脯氨酸有与 α 碳和二级酰胺相连的环状侧链。在肽链中，脯氨酸残基上的氮没有质子化位点和裂解位点，这极大地阻碍了裂解，导致在脯氨酸不能裂解的位点缺少 b 或 y 离子。

8.6 最好的方法

串联质谱现已成为测定肽序列的最好方法。许多年来 Edman 降解作为标准的多肽序列测定方法，但这种方法有很多局限性。第一，Edman 降解不能分析酰胺末端被修饰的肽（Edman 试剂不与 N 端被修饰的肽反应）。串联 MS 不仅能分析 N 端修饰肽，而且能揭示修饰的性质。第二，尽管 Edman 降解可用来找出某些翻译后修饰（如磷酸化），但其通用性远不如质谱方法，因为 Edman 分析对肽 N 端残基进行连续化学切割，然后用层析法鉴定切割衍生物。如果没有各种标准修饰氨基酸，Edman 分析不能确切鉴定被修饰氨基酸。另外某些修饰可能干扰 Edman 试剂与修饰肽之间的反应。

现在可以认为串联 MS 是肽序列分析的最先进方法，通过比对肽序列与数据库中蛋白质序列我们可以鉴定未知蛋白质。然而，实际上仍存在一个问题：分析串联 MS 数据所代表的序列是一项费时费力的工作。下一章将描述已开发的新工具，这些新工具可以解决上述问题，并使串联 MS 数据能实际用于高通量蛋白质鉴定。

推荐读物

Roepstorff，P. and Fohlman，J. （1984） Proposal for a common nomenclature for sequence ions in mass spectra of peptides. *Biomed. Mass Spectrom*. **11**，601－601.

Yates，J. R. （1998） Mass spectrometry and the age of the proteome. *J. Mass. Spectrom*. **33**，1－19.

9 用串联质谱数据进行蛋白质鉴定

9.1 用 ESI 串联 MS 鉴定蛋白质

有两种可利用肽 MS-MS 谱图鉴定蛋白质的方式。第一种是从谱图的从头解释得到肽序列，然后用得到的肽序列通过序列数据库的 BLAST 检索鉴定蛋白质。如果需要处理的谱图数量较少，这种方法是可用的。由于谱图的复杂性不同和分析人员经验上的不同，手工完成一张 MS-MS 谱图的从头解释需要半小时到几天的时间。如前所述，某些谱图没有完整的 b 或 y 离子系列信息，因此不可能从这样的谱图中清楚地解释肽序列，需要分析人员在缺少谱图信息的地方猜测序列。当然，准确的 MS-MS 从头解释需要技巧和经验。这个方法可以很容易地在一两天内给出样品中几个肽的序列，并通过 BLAST 检索鉴定前体蛋白质。这对从 SDS 凝胶条带鉴定一两个蛋白质是完全可行的。

然而蛋白质组学领域要求从 MS-MS 谱图鉴定大量的蛋白质。很清楚，从头序列分析及 BLAST 检索方法不能很快地实现对大规模蛋白质的鉴定。这种方法中的"缓慢步骤"是用以确定序列的 MS-MS 谱图的手工解释。第二种用 MS-MS 数据鉴定蛋白质的方法在这方面有很大的优越性。

第二种蛋白质鉴定方法绕过"缓慢步骤"（手工从头序列解释），运用算法使 MS-MS 谱图数据直接与数据库中肽序列相关联，实际上并不单独解释每一个 MS-MS 谱图。下面将描述这种工具怎样工作。第二种方法非常适合使用基因组序列测定产生的数据库资源，通过肽 MS-MS 谱图与数据库序列的匹配来鉴定蛋白质。使用这种方法的限制因素是 MS-MS 谱图的质量以及数据库的完整性和精确性。

如果得到一个在数据库中存在的肽序列 MS-MS 谱图，正确的算法能够进行匹配。下面讨论的算法可以使 MS-MS 数据与蛋白质序列或与从核苷酸序列（基因组或 EST）推导出的蛋白质序列匹配。人和其他生物基因组序列测定的进展使数据库更精确和完整。确实，在已进行基因组序列测定的生物中，近期将会出现能代表全部基因的完整蛋白质序列数据库。这样一批数据库信息给分析蛋白质组学带来空前的增长动力和可靠性。

9.2 从 ESI 串联 MS 数据鉴定蛋白质的算法和软件工具：Sequest

使 MS-MS 数据与数据库序列匹配以鉴定蛋白质的第一个算法/程序是 1995

年由 John Yates 和 Jimmy Eng 引入的 Sequest。本书在后面也会介绍另外几个类似的软件工具，但这里主要描述具有代表性的一类工具 Sequest。Sequest 这类程序的价值在于能相对快速地将 MS-MS 谱图指定到数据库中特定的肽序列，这使得蛋白质组学分析中的 LC-MS-MS 数据大大减少。然而，应该强调 Sequest 和类似程序本身并不实际进行谱图的从头解释，所以，这些程序的运算结果依赖于所获得的 MS-MS 数据的质量和所用数据库的完整性和准确性。

下面简要概述 Sequest 是怎样工作的。当 MS 仪器获得一个 MS-MS 扫描时，仪器记录 MS-MS 扫描信息和前体离子的 m/z 值。这些数据一起保存。分析完成后，打开 Sequest 程序，选择含有要分析的 MS-MS 扫描数据文件夹，标明对蛋白质样品进行消化所用的酶（如胰蛋白酶），指定用于 MS-MS 分析的是单电荷离子还是双电荷离子，选择要与 MS-MS 数据进行比对的数据库。

一旦程序开始，数据库中所有的蛋白质序列通过指定的酶（如胰蛋白酶）进行虚拟消化，产生用于 MS-MS 扫描比对分析的主目录。然后如下分析每一个 MS-MS 扫描（图 9.1）。

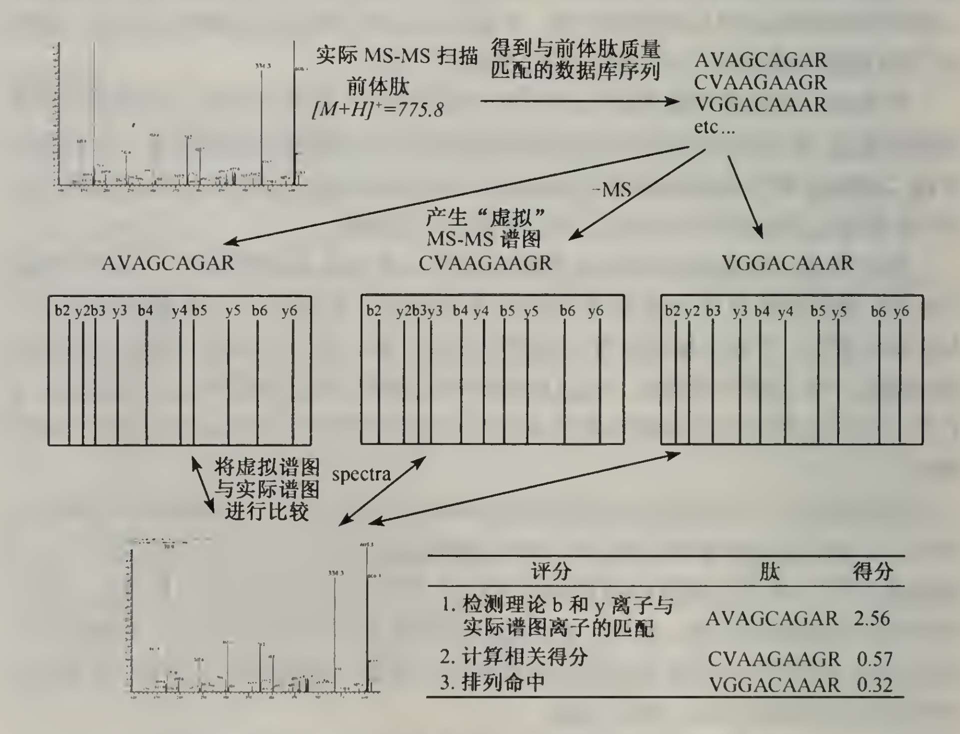


图 9.1　MS-MS 谱图与数据库肽序列相互关联的 Sequest 算法

• 用每一个 MS-MS 扫描的前体 m/z 选择数据库中具有相同质量的肽（在指定的质量误差极限之内）。如果不指定消化酶，程序就选择所有与 MS-MS 扫描分析中肽离子质量相当的肽序列。

• 从每一个选择的肽序列产生理论 MS-MS 谱图。

• 将被分析的 MS-MS 谱图与从数据库产生的所有理论 MS-MS 谱图进行比较。

• 计算在 MS-MS 扫描和理论 MS-MS 谱图之间每一个匹配的相关得分。

然后 Sequest 报告每一个 MS-MS 扫描的最合适的单匹配或多匹配。在一个基于互联网浏览器的视窗中显示在一个数据文件夹中（如一个 LC-MS-MS 实验）的全部 MS-MS 扫描分析的结果。在结果中显示与 MS-MS 谱图匹配的所有特定蛋白质的肽序列（图 9.2）。我们可以根据报告的相关性得分，或通过将“最佳匹配”肽可能的 b 和 y 离子叠在实际 MS-MS 谱图上进行直观检查，来评估一个 MS-MS 谱图与数据库条目匹配的情况，这样可相对容易地区分可靠匹配与不可靠匹配。例如，一个 MS-MS 谱图的主要信号与预测肽大部分 b 离子和 y 离子相匹配常常是正确匹配（图 9.3）。另一方面，大多数主要片段离子与推测肽可能的 b 或 y 离子不匹配的谱图通常是不正确匹配（图 9.4）[①]。

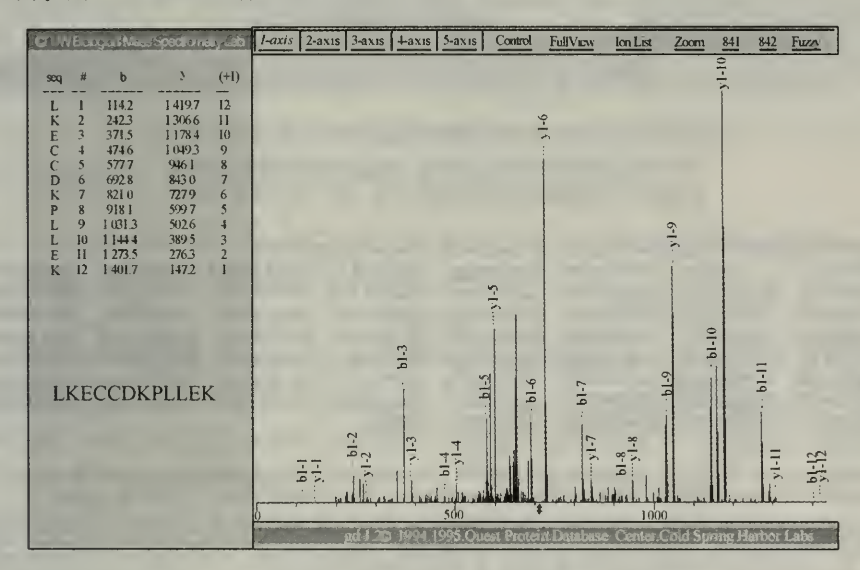


图 9.2　实际 MS-MS 谱图产物离子与匹配肽序列推测的 b 和 y 离子的对应关系的 Sequest 浏览器输出视窗

实际谱图与从匹配肽序列推测的 b 和 y 离子有很好的匹配。

应该认识到 Sequest 不对得到的匹配质量作出判断。即使匹配情况很差，算法也将鉴定出数据库中与每一个 MS-MS 扫描分析数据匹配得最好的肽序列。因而，使用者需有一定的知识、经验和一定程度的直觉来决定哪些匹配是可用的，哪些匹配需要舍弃。在浏览器视窗中显示的与 MS-MS 扫描相匹配的所有数据库

---

① 原文图 9.2、图 9.3、图 9.4 标错。图 9.4 应是图 9.2，图 9.2 应是图 9.3，图 9.3 应是图 9.4，但是译文中未行调整。——译者注

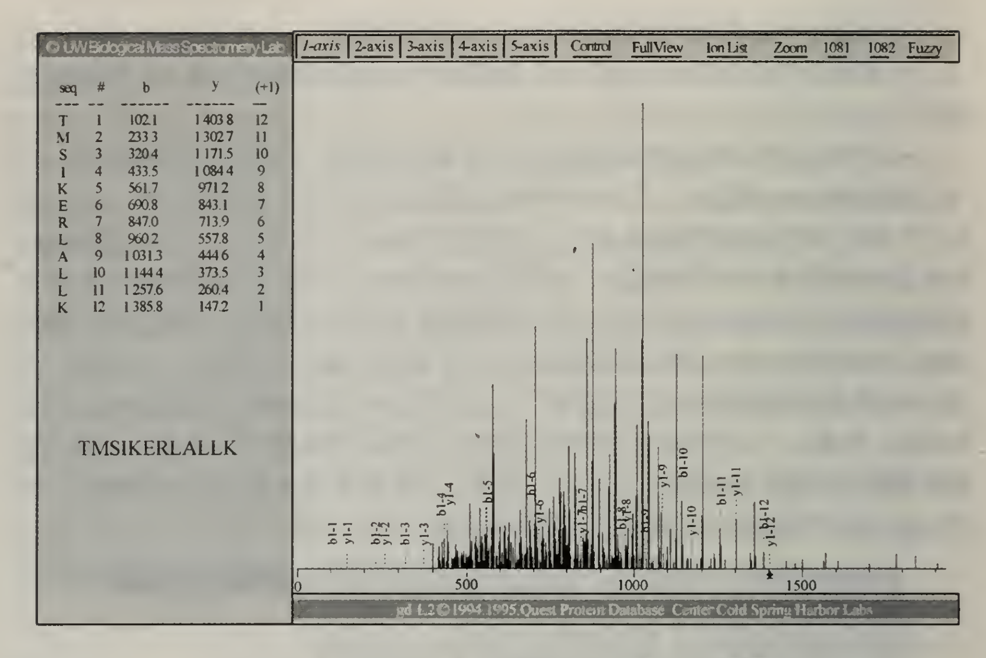


图 9.3　实际 MS-MS 谱图产物离子与匹配肽序列推测的 b 和 y 离子的对应关系的 Sequest 浏览器输出视窗

实际谱图与从匹配肽序列推测的 b 和 y 离子没有很好的匹配。

>gi|418694|pir||ABBOS serum albumin precursor [validated] - bovine [MASS=69270]
MKWVTFISLL LLFSSAYSRG VFRRDTHKSE IAHRFKDLGE EOFKGLVLIA FSOYLOOCPF DEHVKLVNEL TEFAKTCVAD
ESHAGCEKSL HTLFGDELCK VASLRETYGD MADCCEKQEP ERNECFLSHK DDSPDLPKLK PDPNTLCDEF KADEKKFWGK
YLYEIARRHP YFYAPELLYY ANKYNGVFQD CCQAEDKGAC LLPKIETMRE KVLASSARQR LRCASIQKFG ERALKAWSVA
RLSQKFPKAE FVEVTKLVTD LTKVHKECCH GDLLECADDR ADLAKYICDN QDTISSKLKE CCDKPLLEKS HCIAEVEKDA
IPENLPPLTA DFAEDKDVCK NYQEAKDAFL GSFLYEYSRR HPEYAVSVLL RLAKEYEATL EECCAKDDPH ACYSTVFDKL
KHLVDEPQNL IKONCDOFEK LGEYGFQNAL IVRYTRKVPQ VSTPTLVEVS RSLGKVGTRC CTKPESERMP CTEDYLSLIL
NRLCVLHEKT PVSEKVTKCC TESLVNRRPC FSALTPDETY VPKAFDEKLF TFHADICTLP DTEKQIKKQT ALVELLKHKP
KATEEQLKTV MENFVAFVDK CCAADDKEAC FAVEGPKLVV STQTALA

Mass(average): 69270.4 Identifier: gil 418694 Database:C:/Xcalibur/database/bovine.fasta
Protein Coverage:232/607=38.2% by amino acid count,26178.1/69270.4 = 37.8% by mass

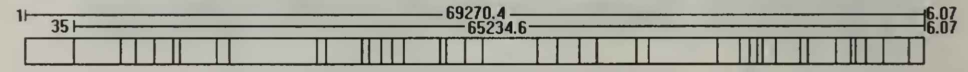


图 9.4　Sequest 浏览器输出视窗

视窗显示了由于 MS-MS 谱图与肽序列的相关性而得到匹配的蛋白质的序列覆盖范围。

蛋白质序列以命中数目（即 MS-MS 扫描匹配）下降的次序列出蛋白质，这将有助于工作人员做出选择。在不同肽序列中有几个高质量命中的蛋白质很可能是鉴定正确的蛋白质。另一方面，与 MS-MS 谱图有一两个弱匹配的蛋白质的鉴定可能不正确。最可靠的蛋白质鉴定是在被鉴定的蛋白质中，有几个不同序列都与数据文件夹中 MS-MS 谱图有高质量的匹配。

有许多复杂情况使 Sequest 分析更耗时或预测不准确、不完全。首先，许多

肽有共价修饰，共价修饰会改变实际分析肽的 m/z 值。Sequest 可能选择数据库中错误的肽进行匹配。在这种情况下，由于这种质量不同，在修饰肽的 MS-MS 扫描和数据库序列之间不会有正确匹配。为了解决这个问题，Sequest 允许使用者指定氨基酸的特定修饰，以便算法能够检索修饰肽和非修饰肽，这很适合可预测的修饰（如丝氨酸、苏氨酸和酪氨酸的磷酸化）。然而，不可预测的修饰也很常见，可能会被遗漏。Sequest 分析的另一个问题是 MS-MS 谱图的前体离子电荷状态（单电荷离子还是双电荷离子）的错误指定。如果一个单电荷离子被错误设定为双电荷离子，它将与数据库中错误肽的理论 MS-MS 谱图比较。双电荷离子如设定为单电荷离子也会产生同样的问题。

虽然一些问题需要引起注意，但不应该分散我们对这样一种极具价值工具的关注。Sequest 分析可以在半小时之内完成含有约 2 000 个 MS-MS 扫描数据的文件，这取决于使用的数据库和计算平台。由 Sequest 提供的蛋白质匹配的质量有时可在数据检查的几分钟内，一般可以在一两个小时之内得以评估。这与要花费成百上千小时进行手工从头解释和对推测的序列进行 BLAST 检索相差悬殊。Sequest 和类似的程序给使用者提供了迅速评估大量 LC-MS-MS 数据的能力。当与自动 LC-MS-MS 仪器控制（如数据依赖扫描）和自动样品制备方法结合时，Sequest 和类似工具可以进行自动和高通量的蛋白质鉴定。

9.3 从 ESI 串联 MS 数据鉴定蛋白质的其他算法和软件工具

其他算法和软件工具也可用来对 MS-MS 谱图数据和肽序列的理论 MS-MS 谱图进行比较。最早开发的 MS-Tag 程序（http：//prospector. ucsf. edu）用来分析肽在 MALDI-TOF 分析中得到的 PSD 谱图（见第 6 章），但现在已被修改用来分析来自不同类型仪器的 MS-MS 数据。程序使用者可以输入待分析的 MS-MS 谱图的 m/z 值目录、前体离子的 m/z 值和电荷状态、用于蛋白水解消化的酶类型和用于得到 MS-MS 数据的仪器信息。算法预先过滤数据库找到与待分析的 MS-MS 谱图中 m/z 相匹配的肽。程序的运算结果提供与待分析 MS-MS 谱图记录的肽离子相匹配的所有肽片段的表格目录。MS-Tag 特别适合含有亚氨离子（表明单独氨基酸存在的低 m/z 片段）的 MALDI-TOF PSD 谱图分析。

Mascot 程序（http：//www. matrixscience. com）使用基于概率的 MOWSE 算法（见第 7 章）、前体肽离子 m/z 信息和 MS-MS 肽片段离子数据鉴定蛋白质。Mascot 实际上是可用于肽质量指纹谱（见第 7 章）和 MS-MS 数据分析的一组程序。在 Mascot 网站下载的转换程序可帮助 LC-MS-MS 数据文件中的多 MS-MS 谱图的自动输入。PepFrag 也是一种类似的软件工具，可以从下面的网址下载（http：//prowl. rockefeller. edu/PROWL/pepfragch. html）。

推荐读物

Clauser, K. R., Baker, P., and Burlingame, A. L. (1999) Role of accurate mass measurement (+/−10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* **71**, 2871−2882.

Yates, J. R., Eng, J. K., and McCormack, A. L. (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* **67**, 3202−3210.

Yates, J. R., Eng, J. K., McCormack, A. L., and Schieltz, D. (1995) Method to correlate tandem mass spectra of modified peptides to amino acid sequences in the protein database. *Anal. Chem.* **67**, 1426−1436.

10 SALSA：一种采集串联MS数据特征的算法

10.1 超出蛋白质鉴定的情况

前一章描述的Sequest和类似的工具根据肽的MS-MS数据，分析“这些肽来自什么蛋白质”的问题。Sequest和类似的程序非常适合于通过肽MS-MS数据鉴定蛋白质。然而，除了简单鉴定样品中存在什么蛋白质，如果我们想进行其他研究，问题就变得有点不同了。考虑下列情形。

• 样品含有多种蛋白质，但是仅希望鉴定其中带有某些特定修饰的蛋白质，如翻译后修饰（如磷酸化），或是药物或其他化学试剂的修饰。

• 鉴定混合物中某些序列相同但其他方面不同的肽。这可能是由于蛋白质存在野生型和突变型。

• 已知或猜测样品中含有一个特定蛋白质，但是推测蛋白质以多种修饰形式存在。希望测定所有这些修饰形式。

如将在下面几章讨论的，在实际蛋白质组分析中常常会遇到这些问题。在每种情况下，我们不是只想鉴定混合物中所有的东西，而是要找出样品中特定组分的信息。我们的目的是从样品的大量MS-MS谱图中鉴定出感兴趣的MS-MS谱图。在第一种情形中，需要寻找表明存在特定功能基团（如磷酸化的氨基酸）的谱图。在第二种和第三种情形中，需要寻找代表特定氨基酸序列基序的b或y离子系列的MS-MS谱图，这些氨基酸序列基序可能存在于几个不同的肽中。使用SALSA（谱图分析的评分算法）算法可以解决以上三种问题。

10.2 SALSA算法

SALSA检测MS-MS谱图特征并根据所显示的谱图特征和在谱图中的强度给谱图评分。SALSA可以检测MS-MS谱图的四种不同类型的特征（图10.1）。第一种是在特定m/z值出现的产物离子。例如肽丢失化学修饰成为带电荷片段，这个带电荷片段然后在MS-MS谱图特定的m/z值处出现，与产生化学修饰丢失的肽本身的m/z无关。第二种是中性丢失。前体离子的一段中性片段丢失，产物离子与前体的电荷状态相同（如一个双电荷离子丢失一个中性片段产生一个双电荷产物离子）。测定的前体离子和产物离子质量的差值等于丢失的中性片段的质量。第三种特征是电荷丢失。多电荷前体离子丢失一段带电荷片段。例如从双电荷前体丢失一个单电荷片段。最普遍的例子是在双电荷肽离子的MS-MS中形成单电荷b和y离子。然而，这个过程也可以是某些化学修饰的显著特征。第四

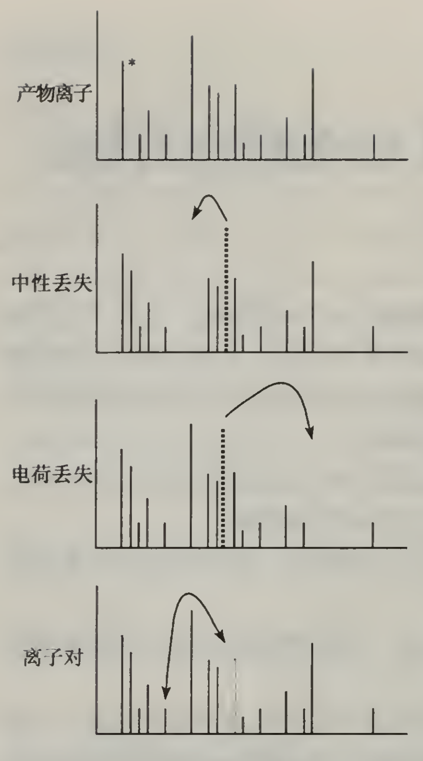


图 10.1　SALSA 算法检测的谱图特征

个特征是离子对，即由 MS-MS 谱图任何位置的一个特定 m/z 值分开的任意两个信号。离子对的出现可以表示在肽序列中存在一个特定组分。例如，肽中含有半胱氨酸残基的 y 系列由于半胱氨酸的质量可能会有一对由 103 m/z 单位隔开的产物离子。

以上 MS-MS 谱图中的一个或几个特征的出现可以成为前体肽中特定结构特征的指标。理论上可以从某些分析中检查单一的 MS-MS 谱图，决定是否存在某些特定特征。但是大量 MS-MS 谱图的解释使得这样做并不现实。我们面临的问题与尝试从复杂肽混合物的 LC 串联 MS 分析中鉴定蛋白质的问题相同。SALSA 算法可满足通过快速的计算机辅助筛选大量 MS-MS 谱图的需要。

当然，仅检测谱图特征是不够的。从大批 MS-MS 扫描数据中鉴定一小部分含有特定特征的 MS-MS 谱图的算法必须能排列出最优的匹配。SALSA 根据指定特定离子的信号强度对 MS-MS 扫描评分。指定的中性丢失（如从磷酸丝氨酸丢失磷酸）产生强信号离子的 MS-MS 扫描得分高。相反，与相同中性丢失相关的较低丰度产物离子的扫描得分偏低。

SALSA 的一个重要特性是其灵活性。我们可以指定检测和评估一些感兴趣的结构特征。如将在后面几章看到的，中性丢失是某些翻译后修饰（如磷酸化）的重要指标，离子对是某些其他修饰的特征（某些稳定氨基酸修饰），产物离子则表明另外一些修饰（某些药物和化学试剂修饰）。

SALSA 评分灵活性的另一个方面是我们可以建立特征重要性的不同等级。某些谱图特征可设定为初级特征，另一些特征设定为二级特征。一旦检测到初级特征，就得以评分。二级特征与某些初级特征相连，只有检测到相关的初级特征，二级特征才得以评分。例如，修饰肽的某些化学组分（如糖）会有水的中性丢失，在某些情况下这可以是特定特征的指标。然而，对显示肽前体有 18 质量单位的中性丢失的 MS-MS 扫描中，其 SALSA 检索很可能有很多命中，因为许多肽离子即使不含有感兴趣的结构特征，也极有可能以这种方式形成片段（如含有丝氨酸或苏氨酸残基的肽离子）。可以将水的中性丢失（18 amu）的得分定为二级特征，以便检测某些更少见的初级特征，如产物离子或其他更少发生的中性

丢失。18 amu 的丢失只有在同时含有其他初级特征的扫描中得分，这两种特征都对扫描的得分起作用。在初级-二级得分等级中，多重得分标准的使用增加了 SALSA 选择性检测肽及其衍生物 MS-MS 扫描的专一性。

10.3 用 SALSA 检索氨基酸序列基序

通过 SALSA 检测 MS-MS 谱图的一个重要特征是在谱图某处出现离子对，这一离子对被 m/z 轴上的某个特定距离分开。MS-MS 谱图中离子对通常的来源之一是 b 和 y 离子系列。例如，第八章讨论的 AVAGCAGAR 肽的 y 系列含有在 m/z 为 477 和 m/z 为 374 的一对离子，它们是 y_5 和 y_4 离子。这两个离子在 m/z 轴上由 103 个 m/z 单位距离隔开，这表明存在一个半胱氨酸残基。与其类似，y_4 和 y_3 离子之间由 71 个 m/z 单位隔开，这相当于一个丙氨酸残基。如果检测由 103 个单位分开的一对离子对的 MS-MS 扫描，SALSA 能够检测到 AVAGCAGAR，也可能检测样品中其他含有半胱氨酸肽的 MS-MS 扫描。如果集中在 y_5 和 y_3 的间距，相当于半胱氨酸和丙氨酸，长度为 174 个 m/z 单位，这可能更具选择性，会挑出含有 CV 或 VC 二肽的 MS-MS 扫描。然而，单个离子对绝不能成为将任意一个 MS-MS 扫描与其他所有 MS-MS 扫描进行区分的很好方式。

寻找某个特定肽的 MS-MS 谱图的最好方式不是仅测定一对离子，而是测定一系列离子。例如，为了找到与部分特定 b 或 y 离子系列匹配的一个离子系列的 MS-MS 谱图，而去检索一个数据文件夹中所有 MS-MS 谱图，就有可能检测到相关肽的 MS-MS 谱图。SALSA 确实可用来检测 MS-MS 谱图中的离子系列。

为了较详细地了解离子系列评分，我们再以肽 AVAGCAGAR 为例子。假定有几百个 MS-MS 扫描数据，其中之一是双电荷离子 AVAGCAGAR。怎样使用离子系列检索找到它？首先我们会想到在大多数肽 MS-MS 谱图中，最强的离子是靠近肽中部切割产生的离子，因此我们的检索应集中在肽中部裂解产生的离子系列。用与三肽序列 GCA 对应的 4 个离子进行检索（图 10.2）。以质量最高的离子作参考，离子系列中第二个离子比第一个离子低 57 个 m/z 单位（甘氨酸残基质量）。第三个离子比第二个离子低 103 个 m/z 单位（半胱氨酸残基质量）。第四个离子比第三个离子低 71 个 m/z 单位（丙氨酸残基质量）。这类离子系列与肽中含有 GCA 基序的 y 离子系列相对应。如图 10.2 所示，指定的离子系列成为带标记刻度的“尺”，可用它来度量数据中的每一个 MS-MS 谱图。在图 10.2A 中，“GCA”尺与 AVAGCAGAR 的 MS-MS 谱图的信号匹配。应该强调的是在离子系列中“GCA”尺只能检测连锁在一起的 GCA，而不能检测 G、C、A 相互分离的离子对。肽中含有甘氨酸、半胱氨酸和丙氨酸，但如果不是 GCA 序列，不会发生匹配。这可在图 10.2B 中得到解释。在图 10.2B 中，肽 AVACAGGAR 的 MS-MS 谱图不与“GCA”尺匹配，尽管这个肽中含有 G、C、A，

在肽的中部有一“CAG”基序而不是“GCA”基序，在 y_3-y_4-y_5-y_6 系列中两个离子（即 y_4 和 y_5）发生改变，以致它们不与“GCA”尺匹配。y_3 和 y_6 离子能够匹配，它们确定了系列的两端，而且重排的基序和起始基序的氨基酸组成相同。

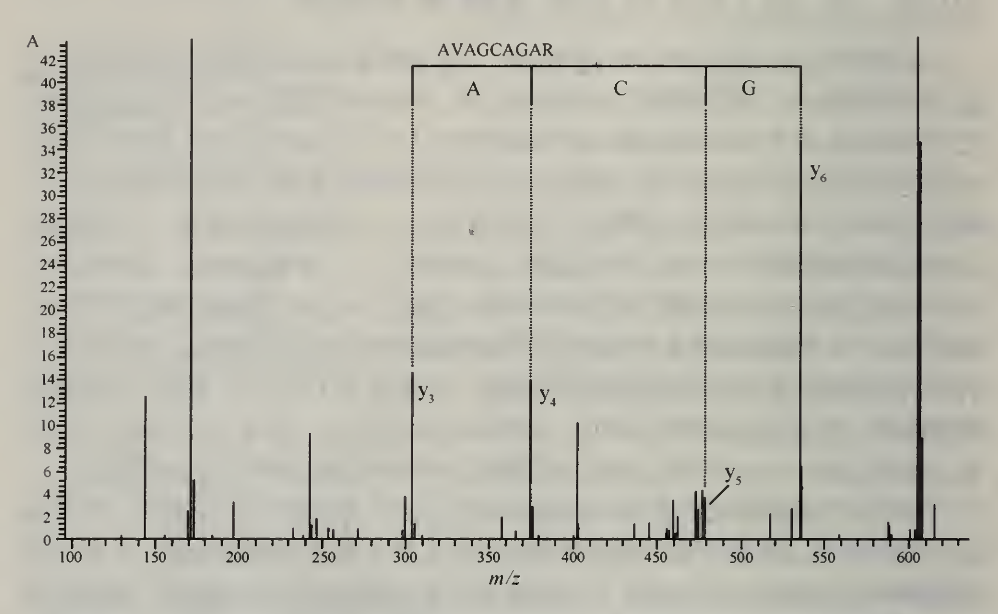


图 10.2A GCA 的 4 离子系列“虚拟尺”与 AVAGCAGAR 的 MS-MS 谱图的匹配

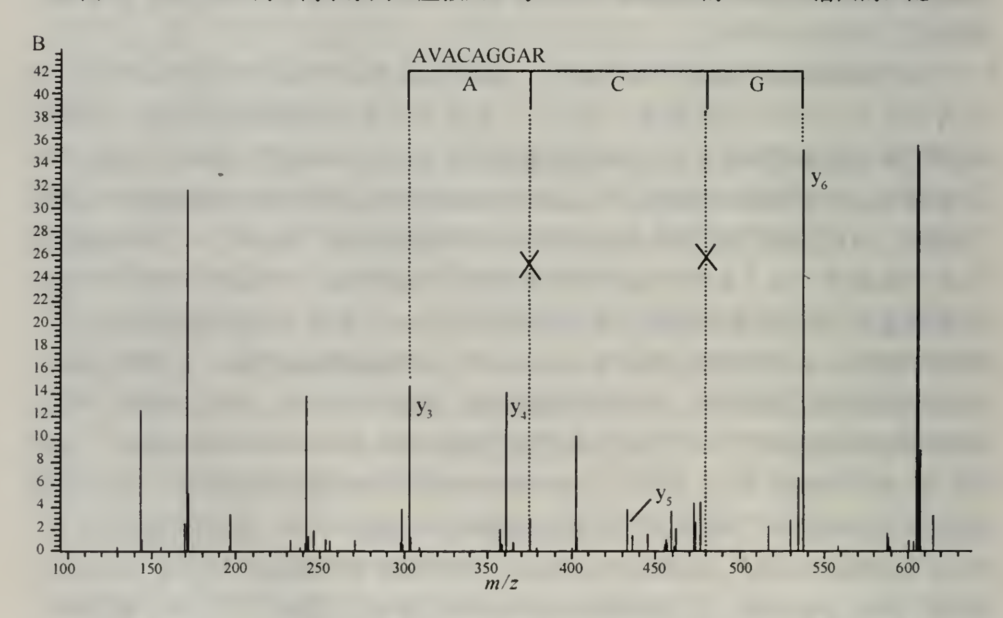


图 10.2B GCA 的 4 离子系列“虚拟尺”与 AVACAGGAR 的 MS-MS 谱图的匹配

用离子系列检索短序列肽的问题是含有 GCA 基序的其他肽也可被检测。而用 VAGCAGA 基序（8 离子系列）的检索极有可能选择性地鉴定出 AVAGCAGAR 的 MS-MS 扫描。

应该明确，SALSA 的一个惯例是从最高 m/z 到最低 m/z 输入一个系列中的离子。用 VAGCAGA 基序的检索从相差 99 个 m/z 单位的最前面两个离子开始（缬氨酸），然后是到相差 71 个 m/z 单位的下一个离子（丙氨酸），之后是到相差 57 个 m/z 单位的再下一个离子（甘氨酸），等等。这个系列的离子相当于 AVAGCAGAR 肽的 y 离子系列（图 10.3）。质量最高离子相当于 m/z 703.82 的 y 离子［VAGCAGAR + H^+］（这个离子不在图 10.3 的谱图中）。系列中的下一个离子相当于 m/z 为 604.69 的［AGCAGAR + H^+］。再下一个离子相当于 m/z 为 533.61 的［GCAGAR + H^+］，等等。在胰蛋白酶肽的 MS-MS 谱图中，y 离子系列常常比 b 离子系列强（这个经验方法也常有例外）。当然，也可以检索相当于 AGACGAV 的离子系列，它会与 AVAGCAGAR 肽的 b 离子系列匹配。

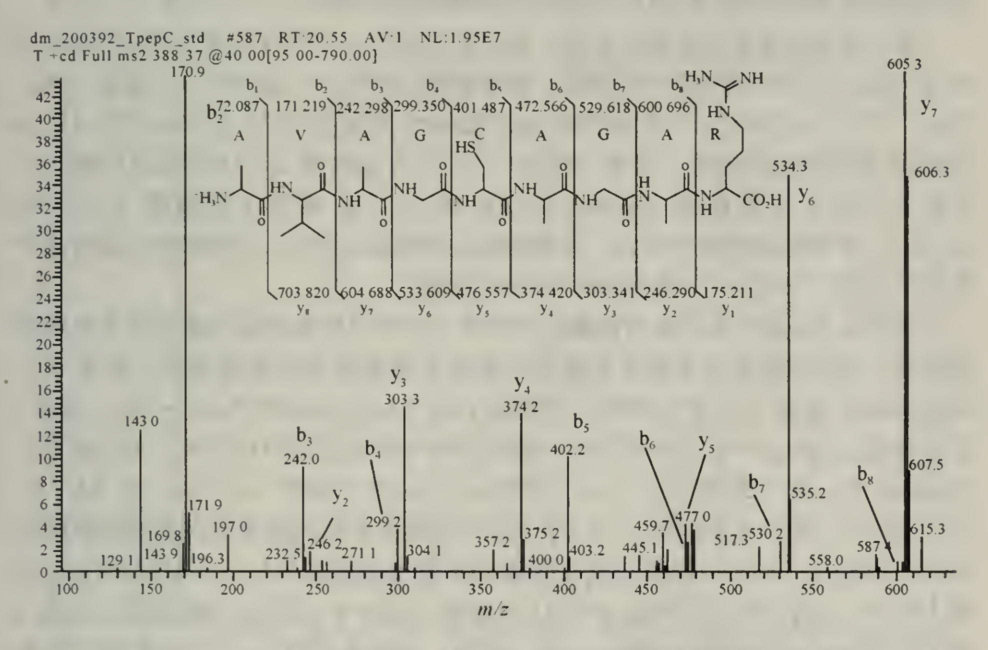


图 10.3 AVAGCAGAR 的 $[M+2H]^{2+}$ 离子的加注释的 MS-MS 谱图

SALSA 根据下列因素对用于离子系列的 MS-MS 谱图评分：①MS-MS 谱图中与离子系列匹配的离子数目；②匹配离子的强度。具有与系列中大多数或全部离子匹配的强信号的 MS-MS 谱图得分最高。另一方面，只与系列中很少几个离子匹配的 MS-MS 谱图，或匹配离子强度很低的 MS-MS 谱图将得到低分。可以

指定在用于匹配的 MS-MS 谱图中必须要有的最少匹配离子数来控制匹配的严格性。

10.4 SALSA 检测离子系列的应用

离子系列检测的应用可以用两个例子来说明。在第一个例子中，我们对检测一个蛋白质的两种形式感兴趣。一个是野生型序列，另一个序列有单个氨基酸发生置换。野生型蛋白质的胰蛋白酶消化产生肽 AVAGCAGAR，而突变型蛋白质用相同的酶消化产生肽 AVAGCAVAR。假定已对其混合肽进行了 LC-串联 MS 分析，希望找出与这两个蛋白质相关的 MS-MS 扫描。序列中相对细微的改变（在第 7 位缬氨酸置换甘氨酸）引起了 y 离子系列的改变。特别是 y 离子系列中 y_3～y_7 由于甘氨酸和缬氨酸的质量不同而发生位移，以致这些 y 离子出现在 AVAGCAVAR 谱图中较高 m/z 位置（图 10.4A 相对图 10.4B）。即使检测的 y 离子出现在不同的绝对 m/z 值，由序列 AGCA 定义的“尺”仍可与图 10.4A 和图 10.4B 谱图中的 y 离子系列匹配。在系列中，y 离子的相对位置在两张谱图中相同。因此对 AGCA 进行 SALSA 检索可以检测这两种肽。

离子系列的检索允许微小差异。如果改用系列 AGCAG 检索 AVAGCAGAR 和 AVAGCAVAR 的 MS-MS 谱图，也能检测这两种肽（图 10.4）。但是“AGCAG”尺与 AVAGCAGAR 的 MS-MS 谱图的匹配要比与 AVAGCAVAR 的 MS-MS 谱图匹配的更好，因为 AGCAG“尺”可以检测 AVAGCAGAR 的 y_2～y_7离子，但是只能检测 AVAGCAVAR 的 y_3～y_7 离子（比较图 10.4A 和 10.4B）。尽管有这种微小差异，检测到的这两种肽的 MS-MS 扫描得分都要高于其他不含有“AGCAG”序列肽的 MS-MS 扫描。

用 SALSA 进行离子系列检索应用的第二个例子是检测蛋白质不完全水解消化产物。蛋白质组学中常会发生蛋白水解酶不能进行完全有效切割。如果 AVAGCAGAR 肽是一个更长序列…FPGKYKAVAGCAGARTGKH…的一部分，不完全消化可以产生 YKAVAGCAGAR 和 AVAGCAGARTGK。YKAVAGCAGAR 的 y 离子系列为 y_1～y_{10}。但其 y_1～y_8 离子系列（y_1～y_8）与 AVAGCAGAR 的 y 离子系列相同，用 VAGCAGA 的离子系列检索可找出这两种肽的 MS-MS 扫描。SALSA 特别有用的是相同检索也能找出 AVAGCAGARTGK 肽的 MS-MS 扫描。因为这个肽含有 C 端延伸（相对于 AVAGCAGAR），它的 y 离子位于 MS-MS 谱图的不同 m/z 值。然而，由检索序列 VAGCAGA 定义的离子系列间隔的距离保持不变，尽管沿 m/z 轴移动。

上述例子再一次表明了用 SALSA 算法进行离子系列检索的能力。在 MS-MS 数据的评估中，图型识别算法可以有选择性地检测与任何肽序列或修饰肽及其不同变化形式相关的 MS-MS 扫描。SALSA 检测能更准确的鉴定肽和源蛋白质。

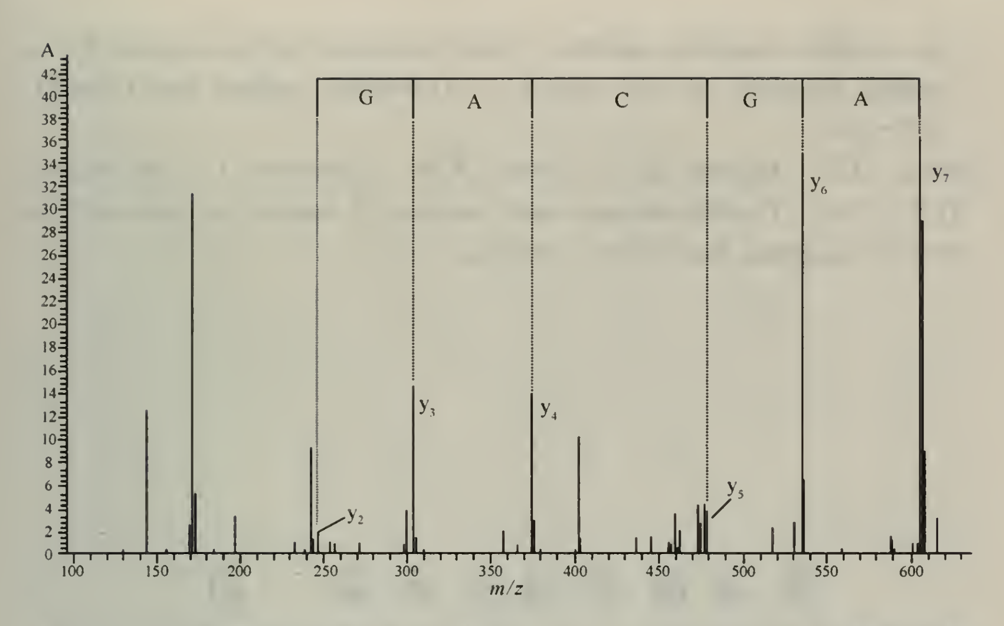


图 10.4A　序列 AGCAG 的离子系列“虚拟尺”与 AVAGCAGAR 的 MS-MS 谱图的匹配

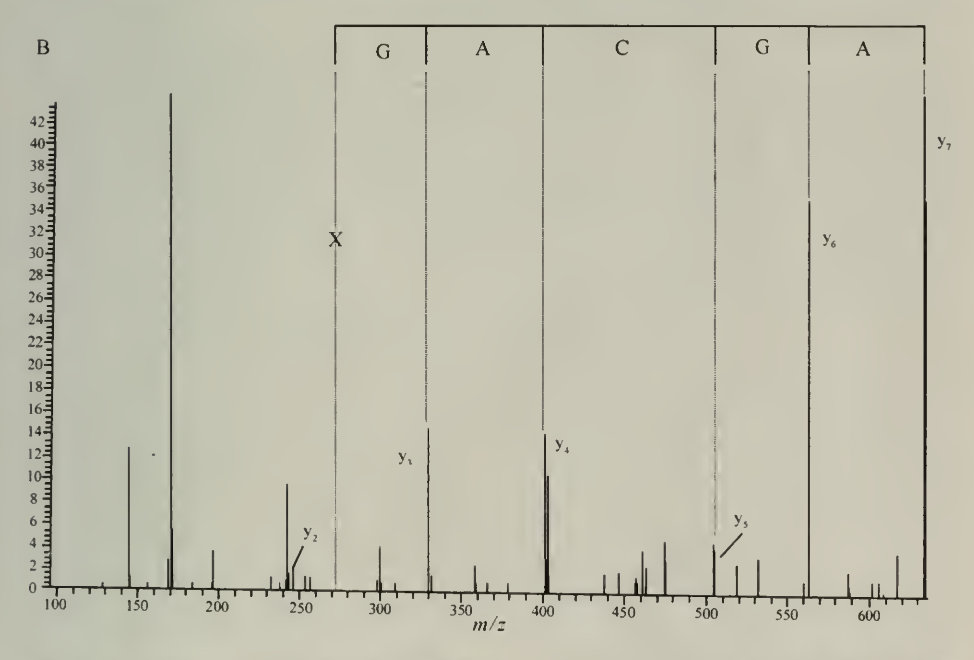


图 10.4B　序列 AGCAG 的离子系列“虚拟尺”与 AVAGCAVAR 的 MS-MS 谱图的匹配

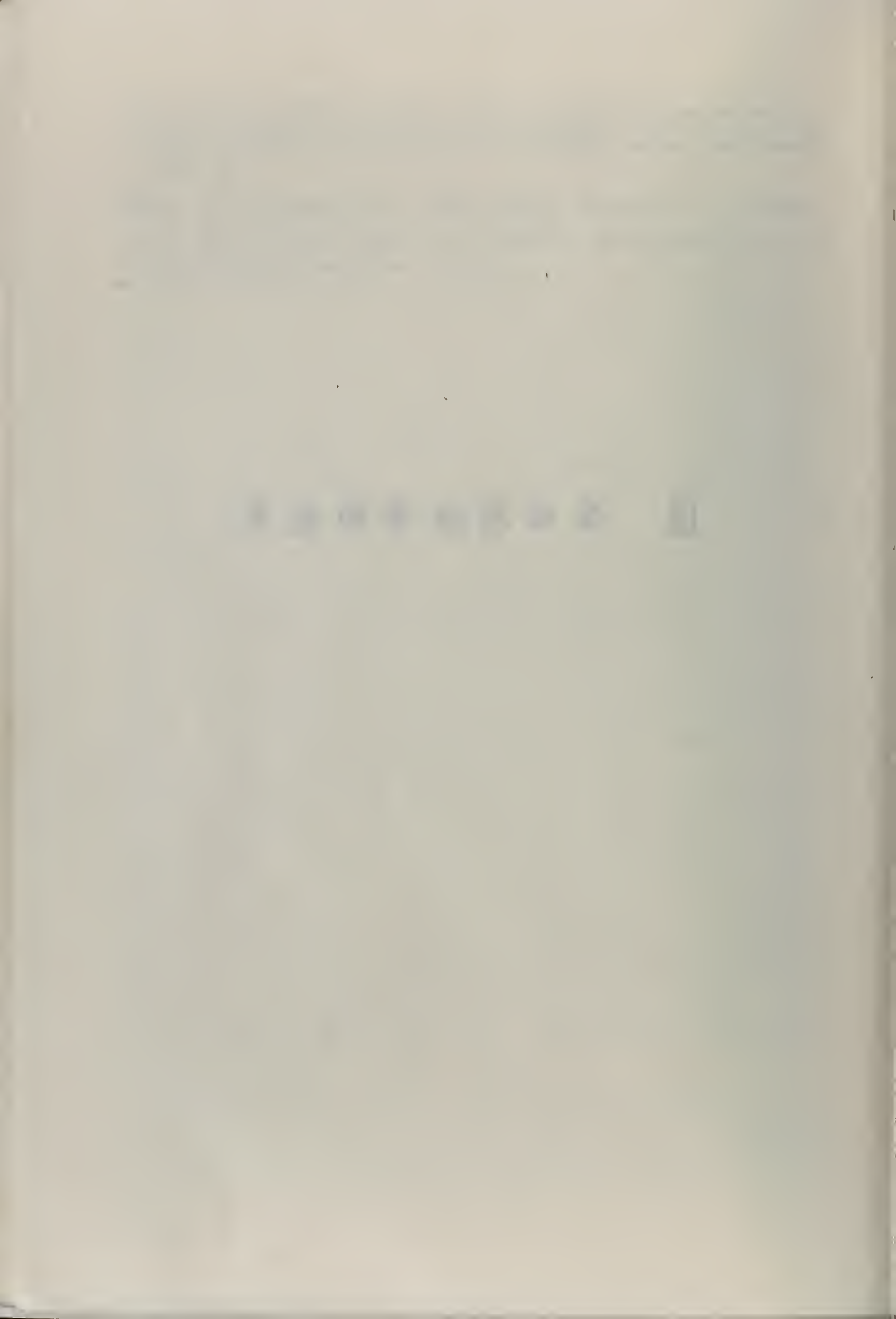
推荐读物

Hansen, B. T., Jones, J. A., Mason, D. E., and Liebler, D. C. (2001) SAL-

SA: a pattern recognition algorithm to detect electrophile-adducted peptides by automated evaluation of CID spectra in LC-MS-MS analyses. *Anal. Chem.* **73**, 1676—1683.

Liebler, D. C., Hansen, B. T., Davey, S. W., Tiscareno, L., and Mason, D. E. (2001) Peptide sequence motif analysis of tandem ms data with the SALSA algorithm. *Anal. Chem.*, in press.

Ⅲ　蛋白质组学的应用



11 采集蛋白质组

11.1 只有一个基因组，但有许多蛋白质组

采集蛋白质组的目的是鉴定尽可能多的蛋白质组分。在前面的讨论中，我们把“蛋白质组”看作是生物所有基因的全部对应蛋白质。然而，“总”蛋白质组是抽象的东西。在任何生物中没有一个细胞同时含有全部基因编码的全部蛋白质。即使在酵母中，在特定时间大约三分之一的基因不表达。

在高等生物中，不同的基因和蛋白质在不同组织和不同的发育阶段表达。例如，眼睛的视网膜色素上皮细胞和主动脉平滑肌细胞中一定有许多相同的蛋白质表达，特别是那些与细胞基本功能相关的蛋白质。然而，表达收缩蛋白质是肌肉的特征，而表达光感受器蛋白质是视网膜色素上皮的特征。在受刺激的肌肉细胞中或在暴露于光中的视网膜色素上皮细胞中可发生更细小的变化，包括蛋白质表达或翻译后修饰的变化。因而不同的细胞表达不同的蛋白质组，相同的细胞在不同状态下表达不同的蛋白质组。

除了生物中不同细胞的许多蛋白质组外，在生物中也有其他有趣的惟一蛋白质组。细胞外液体含有许多分泌蛋白。血浆、CSF、唾液、尿和汗都含有对生物状态变化起反应的蛋白质组。令人感兴趣的是这些蛋白质组分的改变可能与生物的疾病有关，可以用作诊断指标。重要的是任何蛋白质组都是由产生它的生物、组织或细胞的状态来定义。因为状态总是在变化，所以蛋白质组也总是在变化。

11.2 选择用于分析的蛋白质组

不同的细胞和组织含有不同的蛋白质组，因此首先考虑选择什么样的细胞和组织作为蛋白质组样品。要鉴定蛋白质组，我们希望尽可能得到均一的和能代表细胞或组织类型的样品。在某些情况下，培养细胞是相对合适的样品。复杂组织，如哺乳动物的肾，有许多细胞类型紧密连在一起。使组织中的细胞解离和分部的技术可用于复杂组织。然而，在处理样品时靶细胞的生物化学变化可能严重干扰系统。最近为在组织中选择性地收集某些细胞样品而引入的一项技术是激光捕获显微解剖，将显微镜与组织采集结合，从组织的冰冻部分选择性地抽取某些细胞，从而使我们能选择细胞群体进行进一步分析。激光捕获显微解剖在分析肿瘤和相邻正常组织中的应用已为最近癌基因表达变化的研究提供了帮助。捕获用于分析的少量组织或细胞导致只有很少量的总蛋白质可供研究用，因此样品中低丰度表达的蛋白质很难以检测。

从组织或细胞中取样的另一个重要问题是分离步骤往往引起细胞应激反应，蛋白质可能由于对这些应激起反应而发生变化。在细胞取样和 DNA 抽提中形成的活性氧种类（ROS）的增加是我们熟悉的人为假象，这使体内氧化损伤的分析更为复杂。因为 ROS 也能氧化蛋白质并激活适应性响应，所以蛋白质组潜在的人工改变也是一个很重要的问题，特别是在分析与应激有关的蛋白质时这一点更重要。细胞匀浆和分部的第二个主要问题是内源蛋白酶的活化，这可能将导致许多蛋白质被酶解成片段。尽管蛋白质片段的分析仍可产生有用信息，但如果将获得的数据与根据分子质量得到的蛋白质组分相联系，可能会引起混淆。最后。组织样品的制备和储存方式也会极大影响蛋白质组分析的结果。用固定剂固定的组织样品一般不适合于蛋白质组分析，因为用甲醛或类似固定剂处理组织会引起蛋白质交联，妨碍肽片段的消化和回收。相反，冷冻的组织切片很适合蛋白质分析。

尽管蛋白质分部提取可能会对蛋白质组样品有所影响，但对于分析来说蛋白质分部有突出的优点：使分析变得简单。从一个样品检测多种蛋白质的能力取决于从样品中分离到的肽的量和获得 MS 数据的能力。选择分析亚细胞组分（如线粒体）可以使分析人细胞样品的约 25 000 个蛋白质的任务减少到分析 1 000～2 000个蛋白质。用目前的技术，当低丰度蛋白质存在于 1 000 个蛋白质的混合物时，比存在于 15 000 个蛋白质的混合物时更容易得以鉴定。所以目前对细胞或组织样品进行蛋白质组分析的最好方法是通过亚细胞组分分离先对样品进行分部。较简单样品（如 CSF）可能不需要这种预分部。

11.3　第一种蛋白质组采集方法：2D SDS-PAGE 和 MALDI-TOF MS

这种方法将高效的蛋白质分离方法（2D-SDS-PAGE）与方便的、高通量的 MS 分析方法（MALDI-TOF MS）相结合。如图 11.1 所示，蛋白质先从样品中

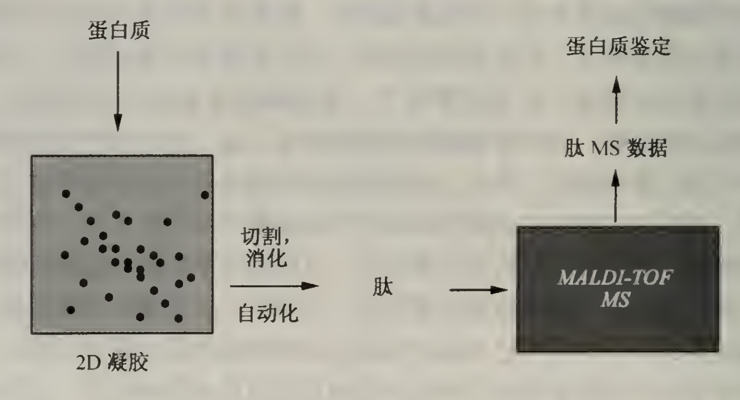


图 11.1　通过 2D-SDS-PAGE 和 MALDI-TOF 进行蛋白质组的采集

抽提，然后用 2D-SDS-PAGE 进行分离。蛋白质点通过染色显示。蛋白质点的手工检查容易引起混淆，可以使用成像系统扫描染色凝胶的图像，记录样品中蛋白质的分布。用几种软件对这些图像进行分析、比较和存档。蛋白质组采集的目的是鉴定尽可能多的系统组分，因此可以选择多个或全部蛋白质点并从凝胶中切出，用凝胶中消化将蛋白质切割成肽，然后用 MALDI-TOF MS 分析（见第 6 章和第 7 章）。在肽质量指纹谱算法和软件（见第 7 章）的帮助下分析 MS 数据，进行蛋白质的鉴定。

如前所述，这个方法的优点是利用 2D 凝胶的分离能力和使用高通量、灵敏和方便的 MALDI-TOF MS 进行蛋白质鉴定。蛋白质组分析一般可鉴定所选择的蛋白质点的 50%～75%，这取决于以下几个因素。

与高等生物相比具有较小蛋白质组的低等生物（如细菌和酵母）一般能得到较大量蛋白质的鉴定。因为这些生物总的基因数较少，在不同基因家族中近乎相同的成员（平行进化同源物）较少（见第 2 章）。相反，高等生物如果蝇和线虫有较大数目的蛋白质平行进化同源物，它们除了有不同的肽，还可能产生若干相同的肽。这降低了根据一两个独特的肽鉴定蛋白质的可信度。

用肽质量指纹谱方法鉴定蛋白质的问题是已发表的基因组序列注释的不确定性。经过测序分析的基因组中的许多基因的编码序列，最初是由帮助定义每一个基因编码序列的起始和终止的算法确定的。进行这些指定的算法有很大的错误比率，可能接近 30%或更多。因而大部分编码序列是正确的，但是某些误差会导致 N 端和 C 端肽的质量指纹谱的错误数据。这个问题会随着基因组序列的更好注释而得以解决，可以在已完成的蛋白质序列数据库帮助下使用质量指纹谱算法。一个相关问题是基因组序列不一定能预测 RNA 剪接变种，而这些变种可能是某些蛋白质的重要形式。在一个蛋白质的两种形式中，会有某些相同的肽。在用肽质量指纹谱算法指定蛋白质时，变种肽会引起不确定性。

肽质量指纹谱鉴定前用到 2D 凝胶，尽管 2D 凝胶的分辨率很高，但仍不能把所有蛋白质完全分离成单个点。由于样品的性质和蛋白质点在凝胶中的位置不同，许多蛋白质点中常含有 2～5 个蛋白质。这些多蛋白质点的消化产生来自 2 个或更多蛋白质的肽混合物，这给肽质量指纹谱软件工具指定相关蛋白质时引入不确定性。尽管较新的算法可以将多蛋白质组分指定到肽混合物的 MALDI-TOF 谱图，但随着混合物复杂性的增加，指定错误也大大增加。

另外，即使采用最灵敏的染色方法，2D 凝胶还是有限定的蛋白质测定动态范围。不同蛋白质的细胞表达水平可有百万倍的差异，而 2D 凝胶蛋白质染色的动态范围最多是百倍到千倍。2D 凝胶一般仅能检测较高丰度表达的蛋白质。用这种方法对酵母蛋白质组进行精细的分析表明 4 000 个表达的蛋白质中只能检测到约 1 500 个，这些蛋白质是高效表达基因的产物。因而，对蛋白质采集来说，2D-SDS-PAGE 和 MALDI-TOF 方法会遗漏大量的蛋白质，包括许多有重要生物

学意义的蛋白质。

11.4 第二种蛋白质组采集方法：多维肽层析和 LC-串联 MS 分析

蛋白质组采集的第二种方法将多维肽分离与串联 MS 肽序列鉴定技术融合在一起。如图 11.2 所示，这种方法中首先将混合物中的蛋白质消化成肽（见第 5 章）。通过层析分离肽（至少部分分离），然后用电喷雾串联 MS 得到肽的 MS-MS 谱图（见第 8 章和第 9 章）。这些谱图在 Sequest 或类似检索工具的帮助下与数据库的蛋白质序列匹配。

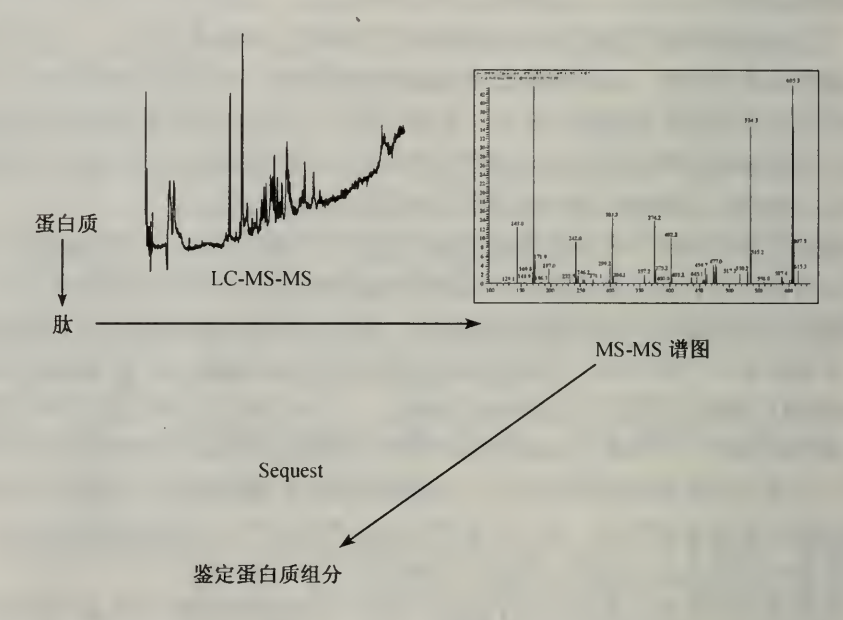


图 11.2 通过 LC-MS-MS 进行蛋白质组的采集

这个基本方法有几个版本，它的两个明显特征是：①主要是用肽而不是用蛋白质进行分析；②是根据 MS-MS 裂解谱图，而不是根据肽质量指纹谱鉴定蛋白质。在这一基本方法中，并不对混合物中的蛋白质进行分离，而是在进行 MS 之前，将蛋白质混合物消化成肽，通过不同的 LC 步骤进行分离。

这个方法的基本思路是得到尽可能多的混合物肽的 MS-MS 谱图。串联 MS 仪器一般使用数据依赖控制，以便选择用于 MS-MS 分析的肽离子并记录数据。为了最多地分析不同的肽，仪器尽可能多地“分散”肽混合物。我们先讨论最简单的 LC-MS 方法，然后考虑几种衍生方法。

1. RP LC-MS。在这一方法中，用 RP LC-MS 分析的肽混合物进行水/乙腈梯度洗脱。通过 RP LC 梯度使混合物中的肽得以分离。这是最简单的系统。这

种系统运转良好，在任何特定时间从柱上可洗脱多于 5 种的肽。但当混合物较复杂时，仪器来不及记录某一时间所有洗脱下来的肽片段信号，在这种情况下，仪器最可能得到在该时间洗脱的丰度最高的 5 种肽离子的 MS-MS 谱图。

2. 具有流动控制的 RP LC-MS。这个方法与前面描述的方法相似，但 MS 仪器能感知“成堆”的肽从柱上洗脱下来。为了避免肽丢失，系统减慢泵流，从而减慢了“成堆”肽离子的洗脱，MS 有更多的时间记录所有肽离子的 MS-MS 数据。特别是当总离子流超过某个阈值时，MS 在 LC 泵上进行反馈控制减慢或停止流动。即使在某一瞬间有几十种肽的复杂混合物进行洗脱，仪器仍有足够的时间系统地得到所有肽的 MS-MS 谱图，而不是只得到最上面 5 或 6 个肽的 MS-MS 谱图。这种策略常称为“峰停留”，可以非常有效地增加鉴定肽组分的数目。为设定反馈控制系统我们需要掌握某些仪器程序编写技巧。这个方法的缺点是停止流动步骤会引起多次停止和起始，导致层析时间延长，产生对计算机辅助数据分析来说数量巨大的数据文件。此外，长时间层析时由于柱上样品的扩散（条带伸展），引起层析分辨率的逐渐降低。

3. 串联 LC-串联 MS（见第 4 章图 4. 7）。在第 4 章描述过这种层析方法，它需要在 RP 柱的前面紧接着装一根强阳离子交换柱。全部样品上样到强阳离子交换柱，大多数肽与柱结合。使用盐梯度洗脱结合不紧密的肽，再用 RP 梯度分离这些肽。盐梯度洗脱后是 RP 分离的循环重复 10～20 次。使用两相层析系统的结果是肽混合物的极大“分散”，同时增加了所鉴定肽的数目。与前面描述的“峰停留”一样，这种方法进行的时间非常长（6～18 小时）并且得到的数据文件非常大。引入这种方法的 Yates 及其同事开发了一种 Sequest 的特定版本，使用群组计算机分析数据。Patterson 及其同事描述的方法中有一个较简单的变换是使用双离子交换方法，RP LC-MS 首先分析的是与离子交换柱不发生结合的肽，然后用单一盐浓度洗脱与柱结合的肽进行第二次 RP LC-MS。这种方法极大增加了不同肽的检测，且不像多步离子交换/RP 循环，只产生较少的数据文件。还有一种版本是使用单独的强阳离子交换柱，用盐梯度将肽分成组，收集到的不同盐梯度组分再依次由前面第 1 部分描述的简单 RP LC-MS 系统进行分析。

4. 有几个“技巧”可以提高前述三种方法的操作。首先应了解只有长度约为 5～20 氨基酸之间的肽能够产生可供数据库检索使用的 MS-MS 信息。因而在 LC-MS 之前可使用大小排阻层析将样品限定为仅含有最可能产生有效鉴定的肽。另一个有趣的想法是使用数据依赖仪器控制系统。可输入一个“排除目录”，这是仪器在运行时可忽略（不进行 MS-MS 分析）的离子目录。样品输入一次，系统将获得其中的多种肽的 MS-MS 数据，然后输入刚刚已分析过的离子目录作为排除目录，再一次分析样品，仪器此时只记录在第一次分析中没有分析过的肽离子数据，这样会增加在两次分析中 MS-MS 谱图记录的总肽离子数目。

前面描述的蛋白质组采集的 LC-串联 MS 方法都开始于消化成肽的粗蛋白质

混合物，然后进行分析。这种方法的特征之一是避免了麻烦的蛋白质分离技术，如 2D 凝胶分离技术。然而，用来自未分部蛋白质混合物的大量肽作为起始材料使肽层析步骤承担了在 MS 分析之前“分散”混合物的重任。可以通过在样品处理前加一个简单蛋白质分离步骤以减少下游分离的任务。有几种简单的分离步骤可用于此目的。通过混合物蛋白质的离子交换层析可将蛋白质分离成 5～20 种组分。类似地，可使用液相制备 IEF。商业上已有的 IEF 装置可产生 12～20 种组分。还可使用 1D-SDS-PAGE。在每种情况下，可容许分离毫克到克的蛋白质。在这些步骤中高容量是重要的，因为这样可增加在 LC-MS 分析中检测低丰度表达蛋白质的可能性。

11.5 哪种方法最好？

前面讨论的蛋白质组采集的两种基本方法有其不同的优点。使用 2D 凝胶和 MALDI-TOF MS 可提供高通量并最大限度地利用强有力的蛋白质分离方法学。使用肽的多维 LC-串联 MS 分析产生最可靠的蛋白质鉴定，因为它是根据直接表明肽序列的 MS-MS 谱图进行鉴定。这两种方法也存在一些问题。2D 凝胶很难重复，2D 凝胶和 MALDI-TOF MS 倾向于检测高丰度蛋白质。另一方面，与串联 MS 一起使用的多维 LC 分离是技术上的挑战，仍在迅速发展。方法的选择取决于已有资源，用任何一种方法都可在蛋白质组采集课题中取得巨大进展。最好是用这两种方法提供互补信息，2D 凝胶和 MALDI-TOF MS 非常适合较简单样品，其目的是鉴定主要系统组分。值得注意的是在 2D-SDS-PAGE 中蛋白质点强度的变化提供了一个明显的比较蛋白质组的方式，这将在下一章进一步讨论。对于复杂混合物中高丰度和低丰度蛋白质的完整采集，很明显与串联 MS 结合的多维 LC 是最有效的方法。最近在酵母蛋白质组采集中，Aebersold 及其同事仔细比较这两种方法就证明了这一点。由于蛋白质组采集的方法正在迅速发展，技术的精确改进和蛋白质分离方法的新组合将进一步“分散”复杂肽混合物，极大增加 LC-MS 方法的效率。

推荐读物

Davis, M. T., Beierle, J., Bures, E. T., McGinley, M. D., Mort, J., Robinson, J. H., Spahr, C. S., Yu, W., Luethy, R., and Patterson, S. D. (2001) Automated LC-LC-MS-MS platform using binary ion-exchange and gradient reversedphase chromatography for improved proteomic analyses. *J. Chromatogr. B. Biomed. Sci. Appl.* **752**, 281－291.

Gygi, S. P., Corthals, G. L., Zhang, Y., Rochon, Y., and Aebersold, R. (2000) Evaluation of two-dimensional gel electrophoresis-based proteome analysis technology. *Proc. Natl. Acad. Sci. USA* **97**, 9390－9395.

Simone, N. L., Paweletz, C. P., Charboneau, L., Petricoin, E. F., and Liotta, L. A. (2000) Laser capture microdissection: beyond functional genomics to proteomics. *Mol. Diagn.* **5**, 301—307.
Washburn, M. P., Wolters, D., and Yates, J. R. (2001) Large-scale analysis of the yeast proteome by multidimensional protein identification technology. *Nat. Biotechnol.* **19**, 242—247.

12 蛋白质表达谱

12.1 处在变化中的蛋白质组

生物化学和生理学的许多工作告诉我们生物化学途径总是在不断变化之中；使用 DNA 微阵证实细胞中基因表达模式也总是在变化的；不同生物技术观测到在生物日常生命周期或在细胞周期中许多酶状态也处于规律性变化之中。所有这些变化提示蛋白质组也处在不断变化之中。除了生命本身的这些变化，环境刺激、化学试剂、药物以及生长和疾病过程都诱导机体产生不同的变化。从事复杂病理学（如癌症）或药物治疗的研究人员对许多诱导变化感兴趣。也许蛋白质组学的最终挑战是测定随时间变化的所有细胞的蛋白质状态，但是技术还不成熟，因此比较一个细胞或生物在两种状态之间的蛋白质组的相关问题变得很重要。

蛋白质表达谱研究中最基本的是测定并比较两个样品中一组蛋白质的表达。这就需要有在两个样品中检测和鉴定相同蛋白质的方法提供比较这些蛋白质水平的基础。以下概述的两种方法可用来比较蛋白质的表达水平，但是这些方法仅表明蛋白质自身的表达水平，不能确定蛋白质修饰的改变。鉴定蛋白质的修饰还需要结合使用后面几章概括的方法。

12.2 使用 2D 凝胶的比较蛋白质组学

比较蛋白质组学最常用的方法是用两个样品进行 2D-SDS-PAGE，比较蛋白质点的图型。2D-SDS-PAGE 特别适合于比较蛋白质组分析，能有效地分离多种蛋白质。最近 2D 凝胶技术的发展（见第 4 章），提高了 2D 凝胶可重复性，甚至在引入 MS 蛋白质鉴定之前，已可以用这种方法进行蛋白质组的比较。蛋白质的鉴定在过去是繁锁和困难的，现在由于肽质量指纹谱和 LC-MS-MS 分析的应用，已可以基本上鉴定任何通过凝胶染色检测到的蛋白质。用 2D 凝胶比较蛋白质组的重要任务是鉴定凝胶之间的不同特征。

研究人员已进行了大量工作来开发分析 2D 凝胶蛋白质点图型的软件工具。此外，也已开发出保存这种信息的大批数据库。2D 凝胶成像分析中最广泛使用的程序是 Melanie™，它是由瑞士生物信息研究所开发的。Melanie™ 对 2D 凝胶染色的图像进行分析。可以使用文件扫描仪（产生 .gif 或 .tif 文件），但最好使用 CCD 相机获得图像。程序首先评估凝胶的“特征”，它是指与凝胶本底相比显现的重要不同。“特征”相当于凝胶上的蛋白质点（图 12.1）。可以通过光密度（OD）、大小和体积（整个蛋白质点面积上的 OD）来鉴定特征。这些鉴定参数是一块凝胶中或多块凝胶之间特征比较的基础。

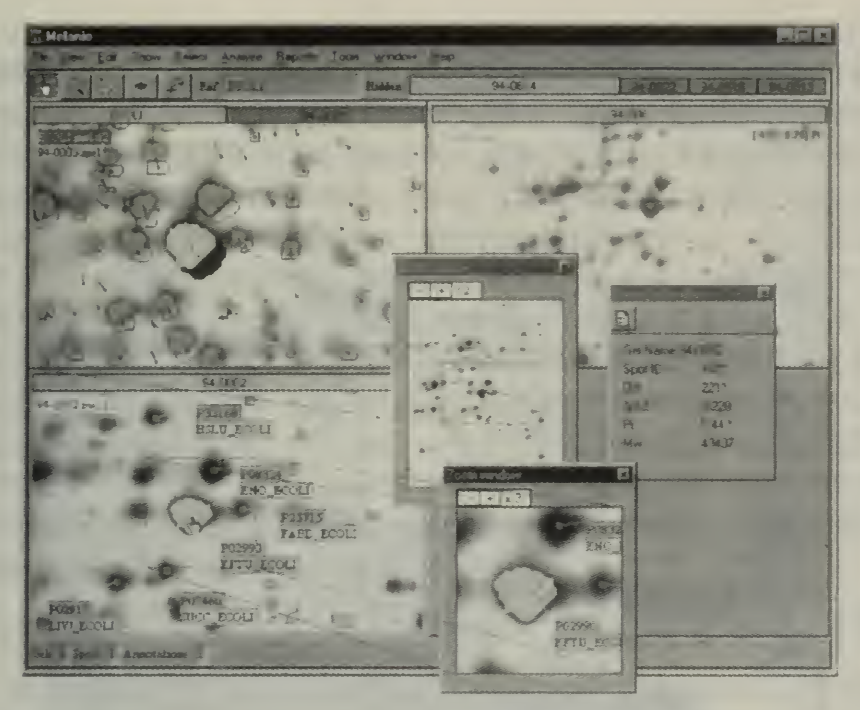


图 12.1　用 Melanie™ 分析软件进行 2D 凝胶蛋白质点图像的注释

当然，对于蛋白质表达谱，我们希望比较两个不同样品在 2D 凝胶上存在的特征点或特征点强度的不同。但是进行多次 2D 凝胶操作时，很难得到精确的重复性。特定蛋白质点的位置通常有微小变化。如果蛋白质点的位置稍有不同，就很难在两块凝胶上比较。为解决这一问题软件允许使用者鉴定“界标”，它们是在要比较的两块（或所有）凝胶中都存在的蛋白质。然后通过软件使这些特征“配对”产生一系列蛋白质对。通过这些蛋白质对可以对比或“匹配”凝胶。匹配过程包括两块凝胶的对比，以便使界标在 2D 空间彼此有对应关系。换句话说，凝胶成像图按像素进行排列以便使所有的界标特征匹配。这个过程需要图像的某些转换或空间“扭曲”，以便补偿凝胶的局部几何变形。

一旦凝胶匹配，就可以进行特征比较，以检测凝胶之间 OD 体积的不同，对可观察到的差异给出图形输出并用数字标出这些区别（图 12.2）。软件也能进行数据统计分析，协助解释一些重要的差异。这种操作允许使用者鉴定两个或更多样品之间的特征或蛋白质点的不同。凝胶可直观地“叠放”以便比较图像。有些软件能从多块凝胶中收集到图像并合成虚拟凝胶，提供在生物不同状态中混合蛋白质组的主档案。

鉴定感兴趣的蛋白质点时包括点的剪切、在凝胶中消化和 MS 分析等操作。除了检测多块凝胶上特征点之间的不同，软件也允许使用者注释特征点，并将特征点与含有 MS 数据、基因序列和功能基因组学数据的数据库文件相连接。

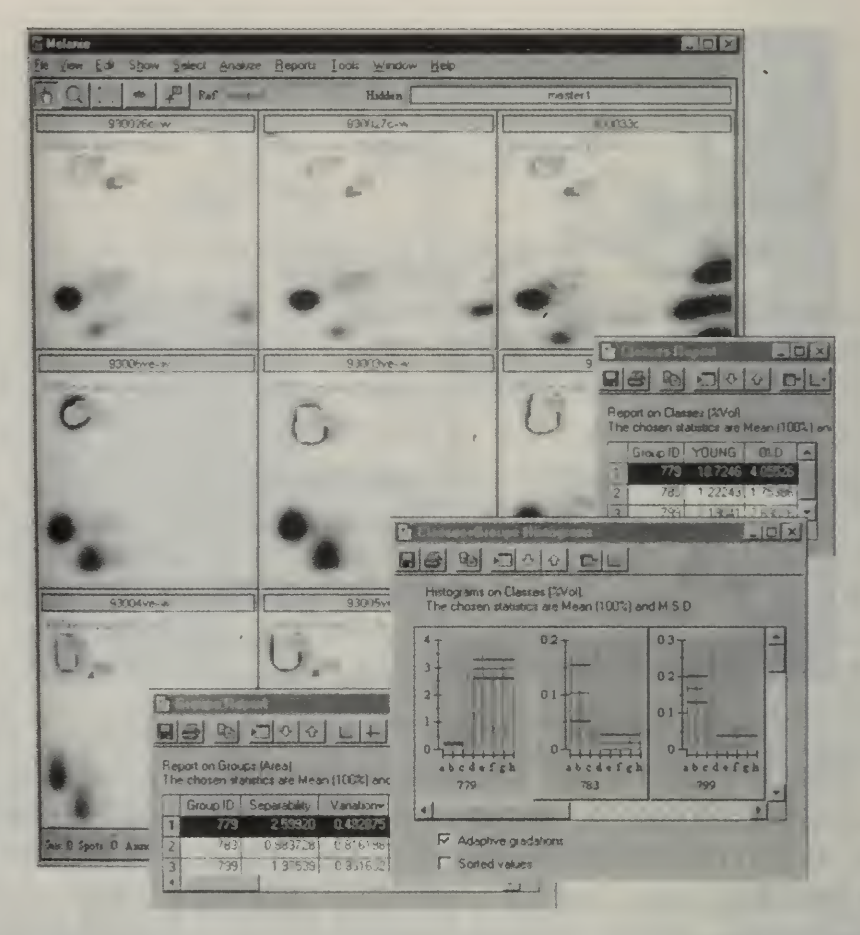


图 12.2 用 Melanie™分析软件对多个 2D 凝胶上蛋白质点的强度进行定量和比较

用 2D 凝胶检测蛋白质表达谱的关键是能有效地比较两块凝胶和鉴定不同表达的蛋白质点，这种方法经过若干研究小组的发展已建立 2D-SDS-PAGE 数据库。这些数据库保存大量经过分析的 2D 凝胶的注释图像。这些数据库越来越成为比较不同实验室所产生数据的重要资源。这些数据库和类似于 Melanie 的软件程序有一个重要特征：能够将大量凝胶与一块凝胶或凝胶组进行比较和进行蛋白质点变化图型的统计总结。

Flicker 程序是另一个通过互联网比较 2D 凝胶图像的独特软件工具，它是由国立癌症研究所的 Lemkin 及其同事开发的（http：//www. lecb. ncifcrf. gov/flicker/)。Flicker 使用与前面描述过的相同方法进行凝胶的评估和匹配。Flicker 的一个重要特征是允许用网上浏览器比较不同来源的凝胶图像。这不仅可以进行来自不同数据库图像的便捷比较，而且可将自己的 2D 凝胶图像与不同数据库的图像进行比较。

2D凝胶是检测蛋白质表达谱的一个重要方法，它提供进行蛋白质组比较的可视图像基础，在这方面是独一无二的。然而，这个方法有一个重要的缺点：2D凝胶的染色仅能测定样品中丰度较高的蛋白质。细胞中蛋白质表达范围约为百万倍，而凝胶染色局限在约一百倍的动态范围。可以通过增加待分析蛋白质的上样量来增加低丰度蛋白质的检测，但是高丰度蛋白质最终控制凝胶的许多特征。许多蛋白质以多种修饰形式存在，它们可能有不同的等电点，可被2D凝胶分开。而对低丰度蛋白质来说，不同修饰的蛋白质分散成多个点可能降低染色检测能力。最后，肽从消化和凝胶中的回收率较低，会妨碍MS鉴定极弱染色（低丰度）的蛋白质。如在第5章提到的，肽从凝胶中消化的回收通常不是定量的，回收率常常少于60%。用2D凝胶检测蛋白质表达谱的主要缺点都来自于2D凝胶染色对蛋白质检测的动态范围有限。尽管染色和显色方法在不断发展和提高，这个问题仍可能最终将2D凝胶局限于分析相对高丰度蛋白质。然而，2D凝胶分析在许多情况下是适宜的，因此基于2D凝胶的蛋白质组谱图将仍是一项有价值和广泛使用的技术。

12.3 使用LC-MS和同位素标记的比较蛋白质组学

对蛋白质组比较来说，LC-MS方法与2D凝胶方法在概念上相反。2D凝胶方法分离蛋白质进行图像比较，而LC-MS方法分离肽，评估样品之间所采集数据的不同。下面是LC-MS方法的概述。用试剂处理两个蛋白质样品以便“标记”。标记物除了一个含有重同位素，一个含有轻同位素，在化学上是相同的。消化样品并用LC-MS-MS分析肽。MS-MS数据的分析（如用Sequest分析）允许鉴定存在的蛋白质。与每一个MS-MS扫描相对应的全扫描谱图可以测定轻同位素和重同位素标记肽的比率。这个比率相当于两个样品中该蛋白质的比率。LC-MS方法不能进行蛋白质的绝对定量，而是提供两个样品中一个特定蛋白质的相对定量。这个方法最早由Gygi和Aebersold用来分析蛋白质组的不同。在下面将讨论这一方法及其某些不同版本。

我们先讨论用于同位素标记的标记物。稳定同位素标记技术是“稳定同位素稀释”技术的改变。要理解为什么使用这项技术，可以回想稳定同位素是原子核中中子数目变化的元素形式，它们不是放射性的。例如，氢没有中子，而它的低丰度同位素氘有一个中子。氢（^{1}H）的质量是1，氘（^{2}H）的质量是2。其他常用的稳定同位素是^{13}C（比^{12}C重1 amu），^{15}N（比^{14}N重1 amu）和^{18}O（比^{16}O重2 amu）。用氘原子置换某些氢原子，这样标记的化合物（称为“含氘的”）有较大的分子质量，因为每一个氘提供额外的1质量单位。一个有8个氘的化合物的质量比未标记的相同化合物质量要高8 amu。然而，这两个化合物至少在我们考虑的分析技术范围内有基本相同的化学性质。这意味着未标记的和用氘试剂标记的同一种肽有相同的层析性质，在MS中显示相同的电

离和裂解。但是 MS 仪器能将它们作为不同的种类加以区分，因为它们有不同的 m/z 值。这些特征使稳定同位素标记成为示踪、标记和在两个不同样品中对相同肽定量的理想试剂。

下面的例子将稳定同位素标记应用到两个不同样品的某个特定肽的分析中。假定样品 A 有 100 pmol 某种肽，样品 B 有 50 pmol 相同的肽。用某个试剂处理样品 A 使其加上一个化学标记 d_0（如在 N 端），而样品 B 带有一个 d_{10}（氘标记）的标记物（图 12.3）。然后混合样品，用 LC-MS 进行分析。因为非同位素 d_0 标记的加尾肽（来自样品 A）和 d_{10}标记的肽（来自样品 B）显示相同的化学性质，经 HPLC 柱一起洗脱，在相同时间进入 ESI 离子源。全扫描谱图（图 12.3）表明两种标记肽的信号。仪器记录 d_0（非标记）和 d_{10}标记肽的单电荷和双电荷离子的全扫描谱图。选择这些肽离子进行 MS-MS 分析。MS-MS 谱图基本上相同（除了预期的由于 d_0和 d_{10}标记质量不同造成的质量不同），这表明在全扫描谱图中两个前体离子代表相同的肽。而且可用 Sequest 分析 MS-MS 谱图找出肽的蛋白质来源。d_0和 d_{10}标记的肽离子强度的比率表明它们在最初样品 A 和样品 B 中的比率。换句话说，d_0标记肽的双电荷离子强度是 d_{10}标记肽的双电荷离子强度的 2 倍，这反映了与样品 B 比较，样品 A 有 2 倍量的特定肽。

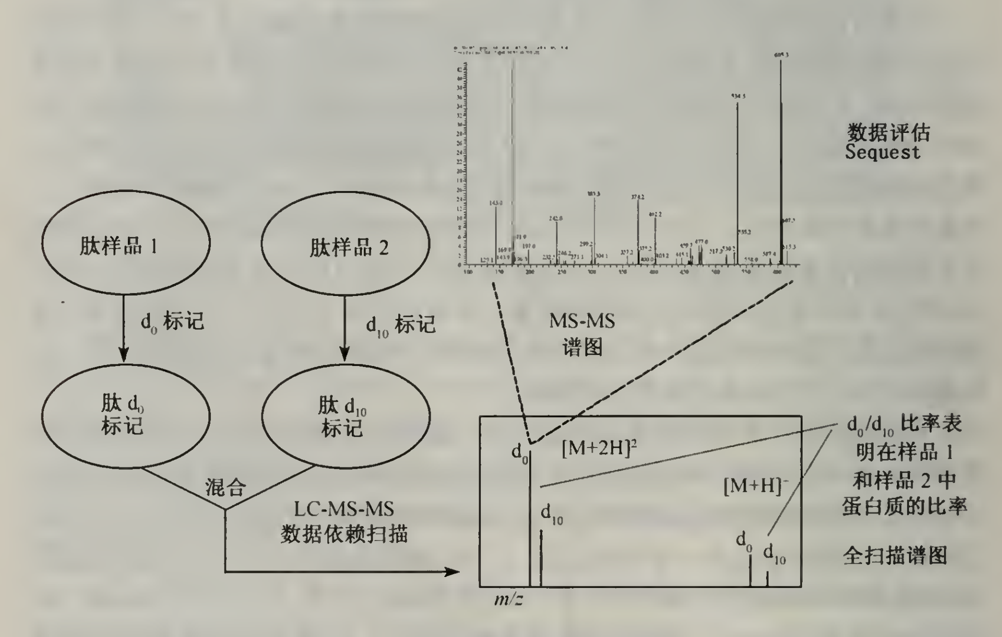


图 12.3　通过稳定同位素标记的 N 端加尾和 LC-MS-MS 分析对两个样品中某个蛋白质相对定量

我们已建立了同位素标记如何通过 LC-MS-MS 帮助肽的定量，再来仔细看一下 Gygi 和 Aebersold 对这个方法的应用。在实际实验中我们常常要对两个样品中多

种蛋白质进行定量比较，而我们的样品多是复杂蛋白质混合物产生的更为复杂的肽片段混合物，每个蛋白质经胰蛋白酶消化产生多种肽，但实际上我们只需要从每一种蛋白质中产生一两种代表肽的标记衍生物以鉴定蛋白质，用以测定蛋白质的相对含量。Gygi 和 Aebersold 的方法使用一种新的多功能标记试剂（图 12.4），称为"同位素编码的亲和标记物"（ICAT）。该试剂由三部分组成，第一部分是和巯基发生反应的碘乙酰胺功能基团，可以使标记物共价标记蛋白质的自由半胱氨酰巯基。第二个部分是含有一个接头，或者含有氢（非标记的，d_0）或者含有氘（标记的，d_8）。第三个部分是生物素，提供对亲和素的高亲和力。

图 12.4　具巯基活性的 ICAT 试剂

图 12.5 总结了分析步骤。蛋白质样品 A 用轻 d_0-ICAT 试剂处理，样品 B 用重 d_8-ICAT 试剂处理。ICAT 试剂标记蛋白质的一个或多个半胱氨酰巯基。样品混合后一起用胰蛋白酶消化，产生含有少许 ICAT 标记肽的复杂消化物。全部混合物上样到亲和素玻珠柱，ICAT 标记的肽通过生物素与柱紧密结合。在混和样品中大多数肽被洗脱掉，保留在亲和素玻珠上的是来自样品 A 和 B 的 ICAT 标记肽。这样，最初非常复杂的胰蛋白酶消化物被简化成一组与亲和素结合的 ICAT 标记肽。这组肽代表其来源蛋白质。然后从柱上洗脱 ICAT 标记肽，用数据依赖扫描通过 LC-MS-MS 进行分析。这些数据的分析基本上与图 12.3 中两种简单肽样品的分析相同。MS-MS 数据的 Sequest 分析（用半胱氨酸残基质量校正存在的 ICAT 标记物）可以找到与其对应的肽和蛋白质序列。因而 ICAT 标记肽产生可以鉴定这些肽的源蛋白质序列信息。与每一个 MS-MS 扫描相对应的全扫描分析揭示了带有 d_0-和 d_8- ICAT 标记物的前体离子。标记肽在整个分析中都成比例，这种比例与标记肽的源蛋白质的比例相符。全扫描中 d_0-与 d_8- ICAT 标记肽之比表明了样品 A 中相关蛋白质与样品 B 中相同蛋白质之比。

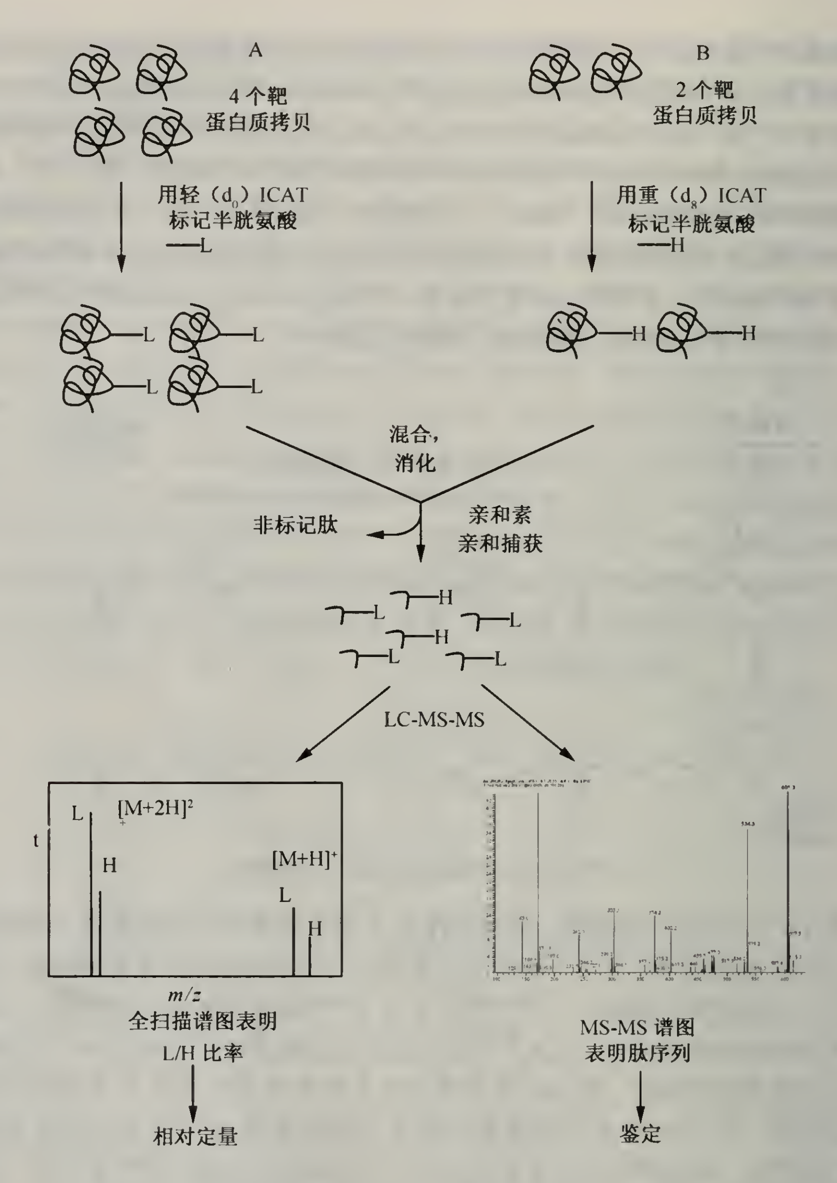


图 12.5　用有巯基活性的 ICAT 试剂和 LC-MS-MS 对两个样品中的蛋白质进行相对定量

例如，酵母抽提物的分析产生两个 ICAT 标记肽 HHIPFYEVDLC* DR 和 DC* VTLK（C* 表示 ICAT 修饰的半胱氨酸残基），这两种肽都定位于蛋白质 GAL10。比较两个样品间 d_0-和 d_8- ICAT-HHIPFYEVDLC* DR 的水平表明两个样品中蛋白质 GAL10 的相对水平，例如比较未经处理酵母和乙醇或半乳糖处理（引起中间代谢中多种酶水平的很大变化）酵母的蛋白质组。分别用半乳糖和乙

醇处理的酵母也可通过 ICAT LC-MS-MS 来比较。乙醇处理的样品用 d_0-ICAT 标记，半乳糖处理的样品用 d_8-ICAT 标记。d_0-与 d_8-ICAT 标记的 HHIP-FYEVDLC* DR 和 d_0-与 d_8- ICAT 标记的 DC* VTLK 的比率是 1∶(>200)，这表明与乙醇处理酵母中的 GAL10 相比，半乳糖处理酵母中的 GAL10 有大于 200 倍的增加。

ICAT 方法一般分析和检测每一个蛋白质的 1～3 个代表肽。检测一个蛋白质的多个肽增加蛋白质鉴定的可靠性。而且多次 d_0/d_8-ICAT 肽比率的测定，可增加测定两个样品中该蛋白质相对含量的精确性。

ICAT 方法在 LC-MS-MS 分析之前使用，使同位素标记与多维肽分离结合。如上一章讨论的，使用与 LC-MS-MS 结合的多维蛋白质和肽的分离，通过“分散”肽混合物以增加相对低丰度蛋白质的检测，使得 MS 仪器能得到样品中尽可能多的肽的 MS-MS 谱图。与同位素标记一起使用多维肽分离可最大程度比较低丰度蛋白质表达的变化。

对蛋白质组比较来说，ICAT 方法有 2D 凝胶/MALDI-TOF 方法不具备的一些优点。第一，LC-MS-MS 比 MALDI-TOF 能更准确地从复杂混合物中鉴定蛋白质。第二，与多维肽分离一起使用的 LC-MS-MS 能检测低丰度蛋白质，而 2D 凝胶分析受蛋白质染色的低动态范围的限制。

ICAT 技术也有某些局限性。第一，某些蛋白质或者不含有半胱氨酸残基，或者半胱氨酸残基在使用的标记条件下不与 ICAT 试剂接触。这些蛋白质则不能用 ICAT 方法检测。第二，ICAT 方法基本上是比较两个样品中蛋白质表达水平的工具。因为在这些分析中只能检测含有可与 ICAT 反应的半胱氨酸的肽，来自蛋白质的大多数肽在亲和素玻珠洗涤步骤被“浪费”。而这些被浪费的肽有许多是关于蛋白质修饰（如磷酸化）变化的信息。这些修饰与两个样品中蛋白质功能的变化有关。除非修饰正好发生在含有可与 ICAT 反应的半胱氨酸的肽，否则不能被检测。

同位素标记方法将被进一步改进。蛋白质组定量比较对理解细胞的生物化学越来越重要。一般方法是在两个样品中用带有不同标记物的标记肽，然后分析样品，比较每个不同标记肽的水平。尽管 ICAT 方法直接标记巯基，它也可能标记肽的其他功能基团，如 N 端胺基。这将改变简化复杂肽混合物的策略，是前述的 ICAT 方法的必要修改。同位素标记策略的创造性应用在定量蛋白质组学中有远大前景。

推荐读物

Binz, P. A., Muller, M., Walther, D., Bienvenut, W. V., Gras, R., Hoogland, C., et al. (1999) A molecular scanner to automate proteomic research and to display proteome images. *Anal. Chem.* **71**, 4981－4988.

Gygi, S. P., Rist, B., Gerber, S. A., Turecek, F., Gelb, M. H., and Aebersold, R. (1999) Quantitative analysis of complex protein mixtures using isotopecoded affinity tags. *Nat. Biotechnol.* **17**, 994—999.

Lemkin, P. F. (1997) Comparing two-dimensional electrophoretic gel images across the Internet. *Electrophoresis* **18**, 461—470.

Wilkins, M. R., Gasteiger, E., Bairoch, A., Sanchez, J. C., Williams, K. L., Appel, R. D., and Hochstrasser, D. F. (1999) Protein identification and analysis tools in the ExPASy server. *Methods Mol. Biol.* **112**, 531—552.

13 鉴定蛋白质-蛋白质相互作用和蛋白质复合物

13.1 “远亲不如近邻”

propinquity 是不常用的名词，它的意思是接触、接近或类似亲属的密切关系。上述引语是证明在取得政治成功中私人接触密切的重要性，是政治上的自明之理。蛋白质在这方面很像政治家，因为在大部分细胞生物化学中像立法政治一样需要团队工作。

蛋白质通过彼此结合形成具有特定功能的多组分复合物，从而使蛋白质“一起工作”。这些功能单位有的如二聚体转录因子复合物一样简单，有的如形成核糖体的 30 多个组分系统那样复杂。生物化学家已开始研究与一个蛋白质结合或相互作用的全部蛋白质。发现在高等生物中（如人和小鼠）蛋白质有多种功能结构域，从而提出许多蛋白质有多种相互作用。理解蛋白质复合物怎样行使功能对理解细胞如何作为一个系统行使系统功能是重要的。

理解这些系统的第一步是鉴定组分。研究者寻找许多新发现基因的功能时的一个重要线索是相关蛋白质与其他已知功能蛋白质的相互作用。例如，许多蛋白激酶信号传导复合物包括激酶、磷酸酶、调节蛋白和结构蛋白的相互作用。尽管可以通过鉴定催化结构域确定激酶和磷酸酶的基本生物化学功能，但辅助蛋白质参与有功能的激酶信号传导复合体的证据主要来自已经证实的这些蛋白质的相互作用。

13.2 鉴定蛋白质-蛋白质相互作用

在细胞系统中蛋白质相互结合主要来自两类实验。第一类实验是靶蛋白质与结合蛋白质一起进行的免疫沉淀（图 13.1）。蛋白质用 1D-SDS-PAGE 分离并电转移到膜上。用推测的可与靶蛋白质结合的蛋白质抗体进行免疫检测。当然，这种方法需要有可与蛋白质结合的抗体，而且往往需要进行推测。这些抗体“沉淀”实验对证实推测的蛋白质-蛋白质相互作用是非常有用的工具，但这种方法无法测定不可预测的多蛋白质复合物成员。例如，图 13.1 中没有对标记为“X”的蛋白质抗体，即使存在相互作用也不能被检测。下面将会讨论这种方法的一些不同版本。这种方法的主要局限性是实验中仅能检测可预测的蛋白质。

第二类主要方法是酵母双杂交系统（图 13.2）。在这种方法中，两个蛋白质之间相互作用的检测是间接的。为编码两个感兴趣蛋白质的（图 13.2 中的 Pr1 和Pr 2)基因分别与一个转录因子的两个不同结构域的编码基因融合，这一对融合

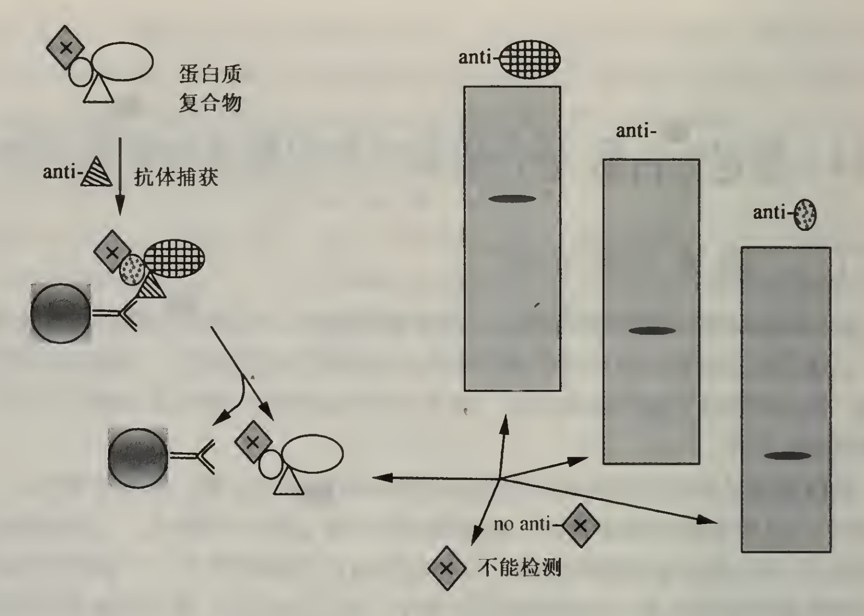


图 13.1　用免疫沉淀和 Western 印迹法对多蛋白质复合物进行解析

的基因在酵母中表达。这两个不同融合体编码的转录因子组分彼此结合［图 13.2 中的 DNA 结合结构域（DBD）和活化结构域（AD）］，活化酵母的报告基因。这种彼此结合是因为我们感兴趣的两个基因产物彼此相互作用形成复合物时才发生。当这两个蛋白质相互作用形成一个复合物时，转录因子的两部分也彼此靠近，使报告基因活化，以便进行信号检测。这是一种重要的测定方法。已经做了很多这方面工作帮助建立来自不同物种的蛋白质-蛋白质相互作用。因为测定方法是间接的，也存在一些可能混淆结果解释的因素，其中包括：①某些融合基

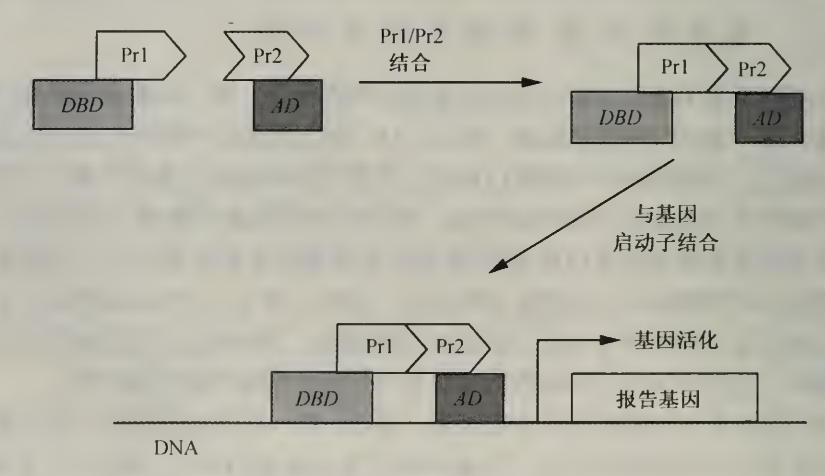


图 13.2　检测蛋白质-蛋白质相互作用的酵母双杂交方法

因产物不能到达细胞核；②融合基因产物与其他蛋白质的相互作用阻碍转录因子组分的活化；③在酵母中某些基因产物很难作为杂合体表达。

13.3 蛋白质-蛋白质相互作用和蛋白质复合物的 MS 分析：基本方法

基于 MS 的蛋白质组分析工具的应用提供了鉴定多蛋白质复合物组分的新方法。基本方法相对简单，如图 13.3 所示。首先我们对某个蛋白质感兴趣（蛋白质 1），与其相互作用的蛋白质未知。制备细胞裂解物，加入蛋白质 1 抗体，蛋白质 1 及其结合蛋白质与抗体发生免疫沉淀。沉淀出的蛋白复合物可以用以下两种方式之一进行分析。

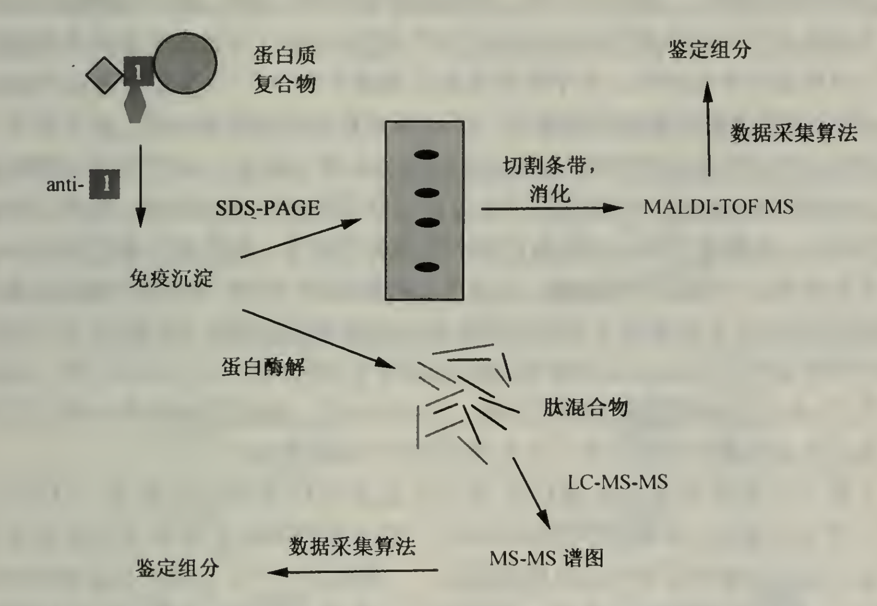


图 13.3 用免疫沉淀和 1D-SDS-PAGE/MALDI-TOF 对多蛋白质复合物进行解析（上面）或用 LC-MS-MS 进行“鸟枪法”鉴定（下面）

一种方法是用 1D-SDS-PAGE 凝胶分离蛋白复合物，然后染色并挑选蛋白质条带进行消化，再用 MALDI-TOF 进行分析。用肽质量指纹谱算法（见第 7 章）从 MS 数据中鉴定蛋白质。这个方法如图 13.3 上半部分所示。也可以在消化凝胶中的蛋白条带后使用 LC-MS-MS 得到 MS-MS 谱图，然后用数据库相关算法如 Sequest 进行鉴定。使用 SDS-PAGE 分离的一个潜在问题是常造成蛋白质的丢失，这是由于在凝胶中蛋白质的不完全消化和肽从凝胶中的不完全回收造成的，这可能会导致混合物中低丰度蛋白质难以检测。另一方面，使用蛋白质分离步骤可以增加用 MALDI-TOF MS 和肽质量指纹谱鉴定蛋白质的效率（见下文）。

另一个鉴定存在于免疫沉淀物中蛋白质的方法是直接对免疫沉淀出的混合物进行消化（不先对它们进行分离），然后用 MALDI-TOF MS 或 LC-MS-MS 分析肽消化混合物。这称为“鸟枪法”分析（DNA 测序策略的类似方法），能得到很好的结果。这个方法如图 13.3 下半部分所示。由于消化物中含有来自复合物中几个不同蛋白质的消化肽段，包括从抗体产生的肽（如下所述，抗体可以从分析中排除），这些因素使肽混合物直接进行 MALDI-TOF MS 分析变得复杂。如果混合物含有三个或更少的蛋白质，用 MALDI-TOF 直接分析鉴定这些蛋白质是相对容易的。随着蛋白质数目的增加，MALDI-TOF 谱图的复杂性使鉴定越来越困难。

这种复杂性使 MALDI-TOF 比 LC-MS-MS 更少地用于肽混合物的鸟枪法分析。LC-MS-MS 更适合于分析较为复杂的肽混合物。LC-MS-MS 鉴定蛋白质是基于用序列相关算法（如 Sequest）分析肽的 MS-MS 谱图。多种不同蛋白质消化产生的肽会产生过于复杂的 MALDI-TOF 谱图，而 LC-MS-MS 获得单独的肽谱图，可单独分析每种肽。对中等复杂蛋白质混合物（2～5 个蛋白质），在 Sequest 帮助下用数据缩减进行简单 LC-MS-MS 可以鉴定结合蛋白质。几个研究小组已使用这种方法鉴定相对较小的多蛋白质复合物。然而，对于更复杂的混合物，会降低每种蛋白质中进行 MS-MS 分析的肽的数目。换句话说，随着许多肽离开 LC 柱，仪器在记录一个肽离子的 MS-MS 谱图时，往往来不及记录另一个肽离子的谱图，导致信号的遗漏，减少了序列覆盖率（即来自每一个被鉴定蛋白质的肽的数目），从而降低了鉴定的准确性。使用多维肽分离（见第 11 章）可以增加序列覆盖率。Link 及其同事使用串联离子交换和 RP LC 与 MS-MS 一起分析了有 78 个亚基的酵母核糖体复合物。与 MS-MS 一起使用多维肽分离大大增加了蛋白质鉴定的数目并增加了每个蛋白质的序列覆盖率。

上面几节表明对免疫沉淀样品的 MS 鉴定可以使用相对简单（MALDI-TOF）或相对复杂（串联 LC-ESI-MS-MS）的方法。方法的选取主要取决于待鉴定蛋白质的量和复合物中蛋白质的数目。一旦建立了 MS 方法并在实验室中运作起来，完成分析是相对容易的。在绘制蛋白质-蛋白质相互作用谱中的真正挑战在于获得发生相互作用的蛋白复合物样品。除免疫沉淀外还有其他分离复合物的方法。下面几节描述分离复合物的方法。

13.4　免疫沉淀

分离多蛋白质复合物最常用的方法是用其中一个组分的抗体免疫沉淀复合物。前面简介的例子使用免疫沉淀分离复合物。我们将讨论成功应用这个方法的某些必要步骤。首先是得到合适的抗体。抗体应该对我们感兴趣的靶蛋白质有专一性结合。此外，抗体不但能免疫沉淀游离的靶蛋白质，而且也能对发生相互结合的蛋白复合物中的靶蛋白质进行免疫沉淀。不是所有的抗体都适合于免疫沉淀，应该对抗体免疫沉淀其靶蛋白质的能力进行测试。

另一个问题是抗体能成功免疫沉淀其靶蛋白质，但是当其他相互作用蛋白质存在时，则不能免疫沉淀靶蛋白质。这可能是由于缺少足够的抗体专一性（即抗体与其他蛋白质反应），或是由于复合物中其他蛋白质与靶蛋白质的抗体识别位点发生结合从而“遮盖”了抗体识别。

抗体对蛋白复合物的分离是有用的工具。然而，当分离到蛋白复合物后，抗体可能使蛋白复合物的分析变得更复杂。与用 MALDI-TOF 或用 LC-MS-MS 分析复合物的肽无关，样品中含有大量从抗体衍生的肽，这些抗体衍生肽使分析变得复杂，降低了鉴定靶蛋白质复合物中成份的灵敏度。防止这一问题发生的最有效方法是使用与不溶性玻珠共价连接的抗体。有一些商业化的试剂盒可以很容易产生固定化抗体。使用这些与玻珠连接的抗体，在免疫沉淀之后用变性剂（如1mol/L 乙酸）简单处理，从与玻珠相连的抗体上释放结合的蛋白质复合体。通过离心或过滤捕获玻珠，回收抗体。洗脱的蛋白质可进行消化和 MS 分析，没有抗体衍生肽的干扰。

一旦鉴定出与靶蛋白质结合的蛋白质，需要用另外的方法证实免疫沉淀实验的发现。我们需要得到某些关于观察到的复合作用不是免疫沉淀实验假象的补充证据。具有已知和相关功能的蛋白质的鉴定，如激酶连同磷酸酶和支架蛋白一起被鉴定肯定是有意义的。然而，我们也可尝试更好地证明靶蛋白质与一种推测的结合蛋白质的结合。通过在另一个免疫沉淀实验中使用对一个结合蛋白质的抗体，可以知道是否起始的靶蛋白质可成为复合物中的结合蛋白质（比较图13.4

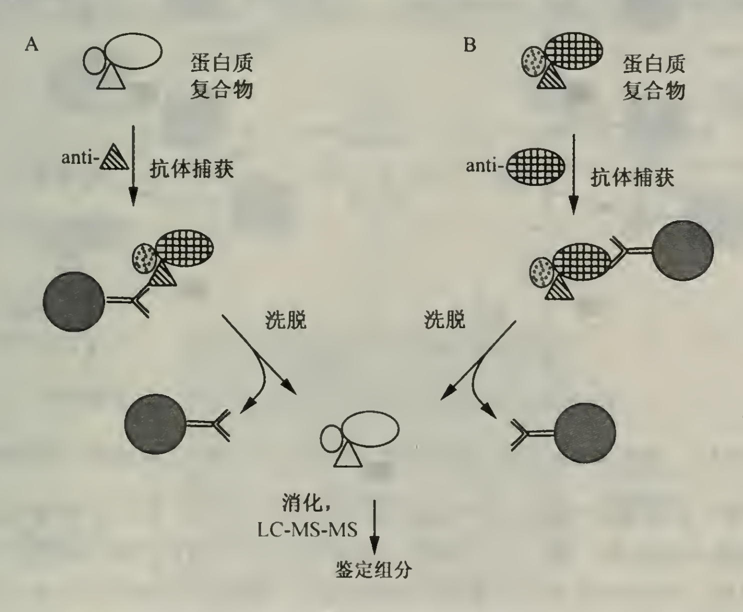


图 13.4　检测蛋白质复合物成员

A. 用对一个复合物成员的抗体进行免疫沉淀；　B. 用对另一个推测的复合物成员的抗体进行免疫沉淀。

中实验 A 和实验 B 的结果)。可以使用一系列这种类型的免疫沉淀证实起始实验中鉴定的蛋白质的结合。

13.5 诱饵和反向诱饵

抗体捕获蛋白质复合物方法的一种变换方式是使用“诱饵”方法。在固相载体上固定靶蛋白质(图 13.5A)。有几种将靶蛋白质固定在玻珠或类似载体上的方法。蛋白质与活化载体如环氧烷琼脂糖凝胶,一起温育可以产生共价连接。环氧烷琼脂糖凝胶与蛋白质亲核的胺或巯基共价反应。另一种方法是产生含有 His 尾部或 FLAG 尾部序列的重组蛋白质。这些重组蛋白质与固定的镍树脂或与固定的抗 FLAG 抗体紧密结合。将玻珠上连接的蛋白质与可能含有结合蛋白质的细胞裂解物或类似抽提物一起温育(图 13.5)。结合蛋白质与固定在玻珠上的靶蛋白质结合形成多蛋白质复合物。通过离心或过滤得到固定化复合物,然后使结合的蛋白质从复合物中解离,再进行消化和分析。选择特定 MS 分析方法的考量与前面描述的分析免疫沉淀蛋白质的考量相同。

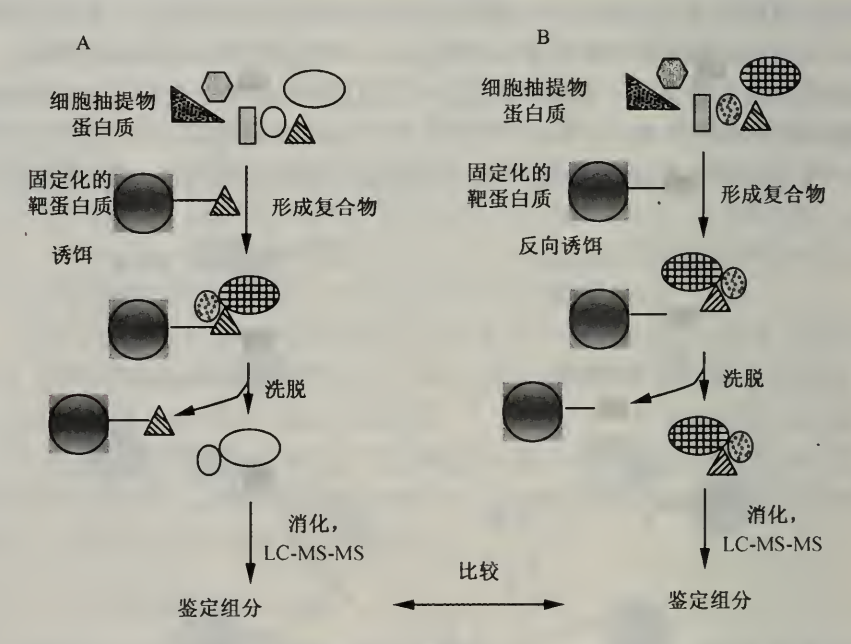


图 13.5 鉴定多蛋白质复合物成员的“诱饵”(A)和“反向诱饵”(B)方法

“诱饵”方法相比免疫沉淀实验有两个优点。第一,不存在捕获复合物中使分析变得复杂的抗体。第二,不需要考虑是否抗体识别复合物中靶蛋白质的适当表位。当然,诱饵方法中也可能由于靶蛋白质与载体的连接而干扰发生在体内的蛋白质相互作用。例如,可能由于靶蛋白质 N 端的固定化破坏靶蛋白质 N 端结构域与另一个蛋白质的相互作用。这个问题可通过下面将要叙述的对照实验适当

避免。诱饵方法相对于免疫沉淀方法的另一个缺点是必须产生或得到固定化的靶蛋白质。取决于靶蛋白质的不同来源，制做固定化靶蛋白比仅仅产生或购买抗体要更昂贵、更困难或耗费更多时间。

使用诱饵方法鉴定了靶蛋白质的结合蛋白质后，可通过“反向诱饵”实验（图 13.5B）证实其相互作用。记录诱饵实验中鉴定出的所有与靶蛋白结合的蛋白质，从所有结合蛋白中选择一个蛋白质进行固定化，然后在相同系统中进行相同类型实验，然后鉴定在“反向诱饵”实验中的结合蛋白质。鉴定出的结合蛋白质中有原来的靶蛋白质和在“诱饵”实验中发现的其他结合蛋白质。这可以证实在该实验条件下这些蛋白质的相互结合。在鉴定出的蛋白质中也可能发现“诱饵”实验中未检测到的结合蛋白质，这可以为在一个网络中寻找其他成员提供新方向。反向诱饵方法的延伸将在这一章后面讨论。

13.6 多蛋白质-核酸复合物

一类重要的蛋白质-蛋白质相互作用是蛋白质与特定核酸序列，如与基因的启动子相互作用。这些相互作用不仅包括与转录和修复相关的几个蛋白质之间的相互作用，还包括蛋白质与特定核苷酸序列的相互作用。鉴定这种相互作用常用的方法是电泳迁移率变动分析（EMSA），如图 13.6 所示。含有感兴趣序列的寡核苷酸探针用^{32}P标记，与含有可能与之发生相互作用的蛋白质裂解物一起温育，抽提后在非变性条件下进行琼脂糖凝胶电泳。不与蛋白质作用形成复合物的寡核苷酸探针在胶中移动较快，而与蛋白质形成复合物的标记寡核苷酸的移动较

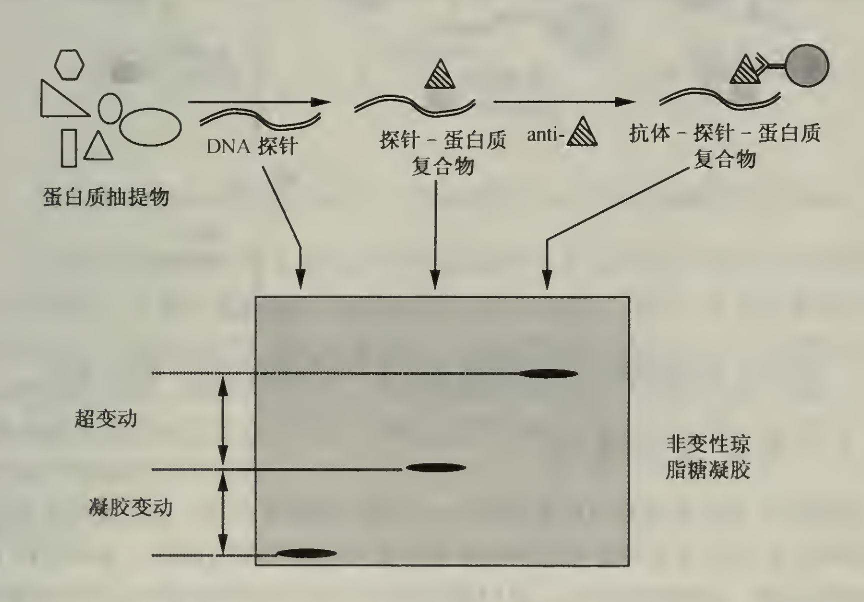


图 13.6　检测与特定 DNA 序列相互作用蛋白质的“凝胶变动”和“超变动”

慢。移动快慢的差别称为凝胶“变动”，这表明（希望如此）一个或多个蛋白质与该序列发生特定结合。电泳之前将推测蛋白质的抗体与复合物一起温育。如果抗体识别并结合复合物中的蛋白质，形成的复合物在凝胶中移动较慢。这种额外的变动称为“超变动”，可提供至少一个复合物成员的鉴定。当然，这种方法的局限性是抗体发生结合并不一定就能鉴定复合物中的蛋白质。更重要的是抗体的作用需要预先推测，这样不利于发现复合物的新蛋白质组分。

分析蛋白质组学可以通过两种方式解决这个问题。第一，切割显示凝胶变动或超变动的条带，消化蛋白质，然后用 MS 进行分析。但是，这些样品中蛋白质含量一般相当小（由于用放射自显影检测^{32}P 的灵敏度），如果不放大实验，或结合使用多个同样的样品，鉴定会较困难。

第二个鉴定与寡核苷酸发生相互作用蛋白质的方法是对前述的诱饵实验稍加改动进行实验。“诱饵”是固定在固相载体上的寡核苷酸（图 13.7）。例如生物素化的寡核苷酸可以用来捕获与核酸相互作用的蛋白质及其结合蛋白质。整个复合物可以用亲和素包被的玻珠捕获，非专一性结合蛋白质可通过洗涤除去，然后洗脱、消化和分析专一性结合蛋白质。这个方法的主要问题是很难捕获与被研究的序列有高度结合专一性的蛋白质。这个方法目前正在几个实验室中开发，最终将在鉴定与核酸相互作用的蛋白质中取代 EMSA 技术。

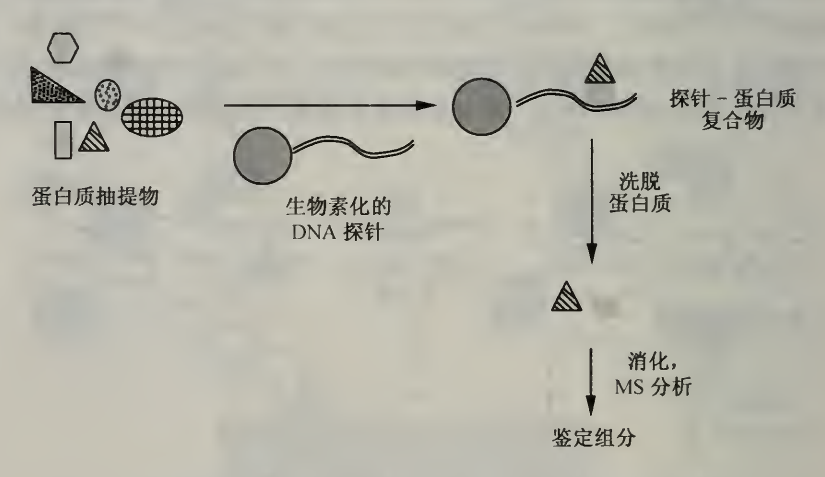


图 13.7　用生物素化 DNA 探针检测与 DNA 相互作用蛋白质的“诱饵”方法

13.7　蛋白质网络谱

细胞中几乎所有的蛋白质都至少与一个蛋白质相互作用，这些蛋白质形成连结的网络。Fields 及其同事在绘制酵母蛋白质相互作用网络谱时，系统应用了酵母双杂交方法。这很好地证实了蛋白质网络的存在。绘制蛋白质相互作用网络谱时对酵母双杂交测定方法稍加改变，改变后的方法是前面描述的诱饵和反向诱饵

实验的扩展应用，这种用于蛋白质网络谱的方法如图 13. 8 所示。首先靶蛋白质，称为蛋白质 1，被固定在玻珠载体上，这个“诱饵”与细胞裂解物温育，回收的蛋白质复合物进行 MS 分析，其结果揭示有几个蛋白质与诱饵结合，它们是蛋白质 2、蛋白质 3 和蛋白质 4。然后用固定化的蛋白质 2 进行反向诱饵实验，该实验表明结合的蛋白质包括蛋白质 1、蛋白质 3、蛋白质 4 和一个新的蛋白质，称为蛋白质 5。下一步制备固定化的蛋白质 5，在相同实验系统中用它作诱饵。与蛋白质 5 结合的蛋白质的 MS 分析揭示了蛋白质 2、蛋白质 3、蛋白质 4 和另一个新蛋白质，称为蛋白质 6 的存在。这种分析循环可以无限延续。对每一个结合蛋白质的发现，可以进行反向诱饵实验证实。

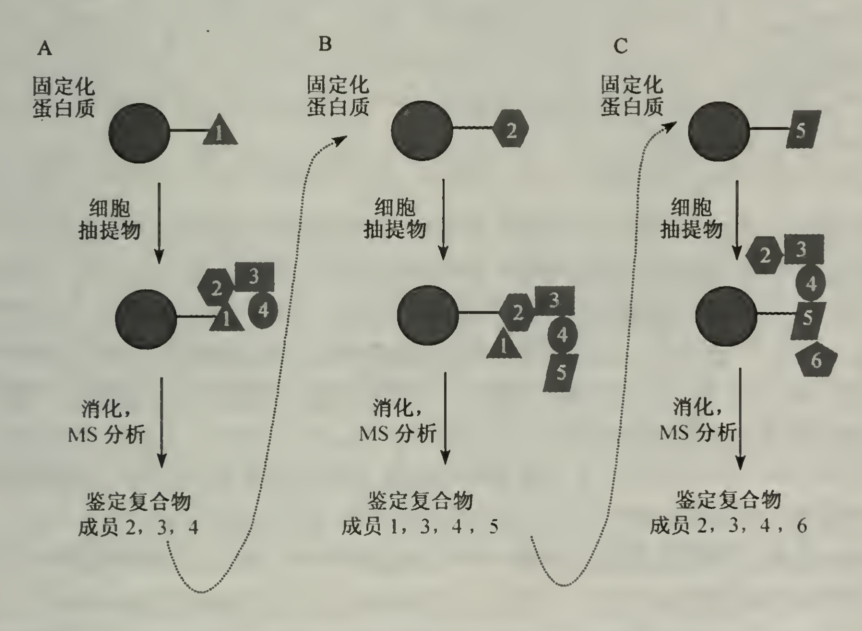


图 13. 8　用一系列连续的“诱饵”和“反向诱饵”实验寻找蛋白质相互作用的网络

这些诱饵和反向诱饵实验的净结果是绘制出蛋白质-蛋白质结合的网络谱，它有可能在一个蛋白质系统中表明这些蛋白质的功能。当然，在这里概括的方法需要在每一个实验中产生固定化蛋白质，这可能是相当耗时的。可以使用抗体进行免疫沉淀实验，总的策略类似。前面描述的使用抗体的优缺点在这也适用。不论是用诱饵还是用抗体进行实验，都在绘制网络谱时确切鉴定每一步相互作用的蛋白质中有其明显的优势。

推荐读物

Craig，T. A.，Benson，L. M.，Tomlinson，A. J.，Veenstra，T. D.，Naylor，S.，and Kumar，R.（1999）Analysis of transcription complexes and effects

of ligands by microelectrospray ionization mass spectrometry. *Nat. Biotechnol.* **17**, 1214—1218.

Honey, S., Schneider, B. L., Schieltz, D. M., Yates, J. R., and Futcher, B. (2001) A novel multiple affinity purification tag and its use in identification of proteins associated with a cyclin-CDK complex. *Nucleic Acids Res.* **29**, E24.

Ng, D. H., Watts, J. D., Aebersold, R., and Johnson, P. (1996) Demonstration of a direct interaction between p56lck and the cytoplasmic domain of CD45 in vitro. *J. Biol. Chem.* **271**, 1295—1300.

Panigrahi, A. K., Gygi, S. P., Ernst, N. L., Igo, R. P., Palazzo, S. S., Schnaufer, A., et al. (2001) Association of two novel proteins, TbMP52 and TbMP48, with the Trypanosoma brucei RNA editing complex. *Mol. Cell Biol.* **21**, 380—389.

Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Seraphin, B. (1999) A generic protein purification method for protein complex characterization and proteome exploration. *Nat. Biotechnol.* **17**, 1030—1032.

Rudiger, A. H., Rudiger, M., Carl, U. D., Chakraborty, T., Roepstorff, P., and Wehland, J. (1999) Affinity mass spectrometry-based approaches for the analysis of protein-protein interaction and complex mixtures of peptide-ligands. *Anal. Biochem.* **275**, 162—170.

Schwikonski, B., Vetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.* **18**, 1257—1261.

Yates, J. R. (2000) Mass spectrometry: from genomics to proteomics. *Trends Genet.* **16**, 5—8.

14 蛋白质修饰谱

14.1 蛋白质修饰无处不在

蛋白质组学显示的一个公理是在生命系统中大多数蛋白质都以多种修饰形式存在。在第 2 章谈到蛋白质的“生命周期”时，讨论了多肽在核糖体的形成、翻译后的切割、通过内源和外源试剂的修饰、氧化损伤的累积和最终的降解，所有这些过程都包括蛋白质结构的不同修饰。某些蛋白质存在多重位点的修饰，这增加了修饰的复杂性。在过去 50 年，生物化学家使用许多蛋白质生物化学技术测定蛋白质的修饰。在大多数情况下，生化测定只能说明蛋白质发生了修饰，但不能表明修饰位点。蛋白质的修饰位点很重要，因为不同位点的相同修饰可能有不同的结果。

在生物化学和细胞生物学最近的工作中，研究者利用许多技术来绘制蛋白质修饰谱。两种常用的技术是抗体和定点突变。我们以蛋白质磷酸化位点的定位为例来说明。针对磷酸丝氨酸、磷酸苏氨酸或磷酸酪氨酸的单克隆抗体（MAb）可以用来定位这些氨基酸残基在完整蛋白质或切割肽段上的位置。定点突变允许“敲除”在研究系统中磷酸化的丝氨酸、苏氨酸或酪氨酸残基。用抗体对这些蛋白质或消化后的肽片段进行 Western 印迹法分析，证明特定氨基酸置换后不再发生磷酸化，从而可以推断哪些氨基酸位于修饰位点。

这一方法有两个问题。第一，不能肯定在定点突变中使用的置换氨基酸是否改变了系统的其他的功能，如与激酶的结合。即使氨基酸置换引起的微小结构变化也可能影响邻近磷酸受体位点的磷酸化。有多个紧密相邻的磷酸化靶点的肽中特别容易发生这种现象。第二个问题是操作问题。这种研究需要大量的工作产生抗体和突变蛋白质。每研究一个新系统时都要解决抗体专一性问题，这极大阻碍了研究进展速度。尽管已用这种方法对磷酸化做了广泛研究，但是有许多有意义的蛋白质修饰不宜采用这种实验方法。

MS 方法的引入为鉴定蛋白质的修饰提供了一种好方法。MS 数据的特点是既可以测定肽质量和序列，也可以提供关于肽修饰的信息。在前面几章讨论的某些 MS 数据采集和软件工具可帮助鉴定特定修饰。这一章分析怎样结合使用 MS 分析和数据采集，在氨基酸水平上准确地绘制蛋白质修饰谱。先以蛋白质磷酸化位点的定位为例，然后延伸到寻找其他内源修饰和由环境因子引起的外源修饰上。

14.2 序列覆盖率是鉴定蛋白质修饰的关键

直到现在我们对蛋白质的 MS 分析主要集中在蛋白质的鉴定和相对定量。在

许多情况下，胰蛋白酶消化后进行 MS 分析可以提供蛋白质多个肽片段的数据。MS 数据在多大程度上代表整个蛋白质序列常称为“覆盖率”。例如，如果分析一个长为 100 个氨基酸的蛋白质胰蛋白酶消化物，得到相关的 60 氨基酸残基的 MS 数据，这被说成是 60%的序列覆盖率。目的只为鉴定蛋白质，这是绰绰有余的。我们在用 MALDI-TOF 数据进行肽质量指纹谱时，只要有 2～3 个肽与数据库条目匹配就足以鉴定蛋白质。实际上，这可以转换成少至 10%～15%的序列覆盖率。在 LC-MS-MS 中，对一个长为 100 个氨基酸的蛋白质来说，如果有两个较好的 6 肽 MS-MS 谱图就可以进行蛋白质的确切鉴定，这仅相当于 12%的覆盖率。因而较低的覆盖率不妨碍蛋白质的准确鉴定。对蛋白质表达的定量也是一样，用 2D 凝胶的图像比较或用稳定同位素标记的 LC-MS-MS，少至 5%～10%的序列覆盖率就可以确切地定量蛋白质的表达特性。

用 MS 绘制蛋白质修饰谱时情况完全不同。只有获得修饰肽的 MS 数据才能检测到肽修饰。要检查一个蛋白质的所有氨基酸的修饰，就必须有全部肽的 MS 数据，即必须有 100%序列覆盖率。这一点如图 14.1 所示。如果进行蛋白质的胰蛋白酶消化，然后得到 MS-MS 谱图和肽 1、肽 4 和肽 7 的序列，可以确切地鉴定蛋白质（图 14.1A）。然而磷酸化发生在肽 2 和肽 8，仅用肽 1、肽 4 和肽 7 的数据不能找到存在于其他肽的磷酸化位点。如果得到含有磷酸丝氨酸残基的肽 2 和肽 8 的 MS-MS 谱图（图 14.1B），不仅能鉴定蛋白质，而且能推测这两个肽的精确磷酸化位点（见下文）。但我们在实际实验中不可能在研究每个蛋白质时都正好获得含有修饰的肽谱图，这个例子表明了在绘制蛋白质修饰谱时覆盖率的重要性。我们必须依赖于覆盖率或运气。但不能总是依靠运气，所以必须提高覆盖率。在讨论提高覆盖率的策略之前，先讨论一下在 MS 数据中什么信息可以帮助我们找到特定氨基酸残基的修饰。

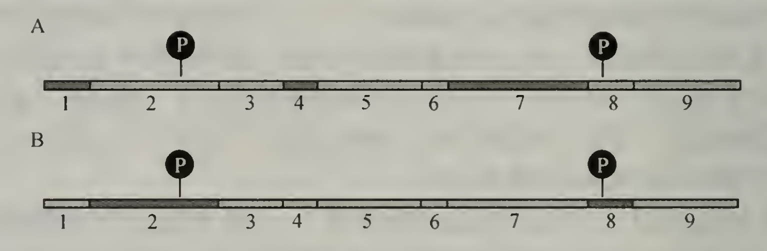


图 14.1　序列覆盖率对检测翻译后修饰的影响

蛋白质的黑色部分代表已得到 MS 数据的序列。肽 2 和肽 8 表明磷酸化位点。

14.3　从 MS 数据推测修饰

使用 MALDI-TOF 得到肽混合物的 MS 谱图，其数据可提供肽离子的精确质量测定。测定的质量反映了肽的组成氨基酸和所有修饰基团的质量之和。从

MALDI-TOF MS 分析我们可以明确哪个肽可能有修饰形式存在。例如，MALDI-TOF 分析磷酸化的和非磷酸化的肽产生两个信号。低 m/z 值的信号是非磷酸化肽，而 m/z 值高出 80 单位的信号相当于磷酸化的肽。

MALDI-TOF MS 分析和肽质量指纹谱算法及软件的结合不仅可以鉴定蛋白质，而且可以鉴定修饰形式。这些软件允许使用者指定如磷酸化的一般修饰以及使用者定义的特有修饰。与数据库中非修饰肽不匹配的信号可能与这些肽的修饰形式匹配。

这对确定特定肽的修饰是有用的。然而，这种方法不能精确地定位特定氨基酸的修饰。例如，肽 VPQLEIVPNpSAEER 仅在丝氨酸残基含有一个可能的磷酸化位点。有些肽可能含有多个可能的磷酸化位点。人 p53 蛋白的肽 GQSTSRHK 含有两个丝氨酸，它们都是激酶磷酸化的位点。非修饰肽有一个 m/z 为 900.9760 的 $[M+H]^+$ 离子，而单磷酸化形式有一个 m/z 为 980.9558 的 $[M+H]^+$ 离子。但是通过这个数值差异我们仍不能确定两个丝氨酸中哪一个发生磷酸化。如果知道蛋白质磷酸化激酶优先磷酸化的序列，能指定可能的磷酸化位点。但这样做也只是进行推断，而不是确切鉴定。要确切鉴定必须获得肽离子的 MS-MS 谱图。

从 LC-MS-MS 提供的肽 MS-MS 谱图，我们不仅能得到序列信息，而且能得到修饰的序列位点。例如，牛酪蛋白的磷酸肽 VPQLEIVPNpSAEER 的 MS-MS 谱图如图 14.2 所示。双电荷离子的质量（m/z 831.2）比预期的非修饰肽的质量（m/z 791.4）高 40 单位。这证实存在磷酸化。谱图含有提供序列信息的 b 和 y 离子系列。谱图显示从 y_5 离子开始 y 离子系列的变化。y_5 离子高于预期的相应非磷酸化肽 80 m/z 单位（磷酸化的肽为 m/z 671.2，而非磷酸化的肽为 m/z 591.6）。而且 y_7、y_8、y_9 和 y_{11} 离子的信号出现在高于非修饰肽相应产物离子 80 m/z 的位置（y_6、y_{10}、y_{12} 和 y_{13} 离子不在谱图中）。仅有一个含有磷酸丝氨酸残基的 b 离子（b_{13}）出现在谱图中，但它的 m/z 移动了 80 amu，这反映了磷酸化的存在。b 和 y 离子系列的改变证实修饰残基的序列位置。MS-MS 谱图（图 14.2）的另一个有意义的特征是位于 m/z 782.1 的强产物离子，这个离子来自中性片段中丝氨酸的磷酸（98 Da）丢失。[一般磷酸（98 Da）从双电荷离子的中性丢失产生一个比双电荷前体的 m/z 低 49 单位的信号，在图 14.2 中，m/z 为 831.2 的双电荷磷酸肽中磷酸的丢失导致在 m/z 为 782.1 处产生一个片段。从单电荷前体丢失相同片段产生一个比单电荷前体 m/z 低 98 单位的信号。] 这种容易发生的丢失是 MS-MS 中磷酸丝氨酸和磷酸苏氨酸残基的特征。相反，磷酸酪氨酸残基不易丢失磷酸，因为它们没有 α 氢帮助磷酸的消除反应。$[M+2H]^{2+}$ 前体离子质量的改变、双电荷离子的 49 单位（磷酸）的中性丢失和含有磷酸丝氨酸的 b 和 y 离子的 80 amu 移动，综合这些因素可以确切证实在这个肽中丝氨酸磷酸化的存在和序列位置。

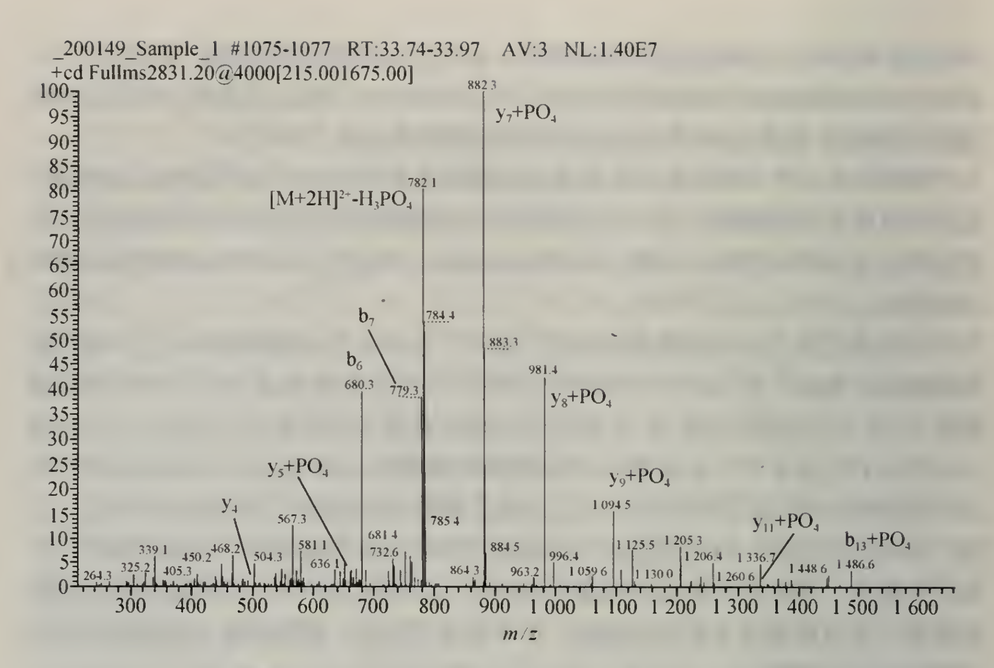


图 14.2 来自牛酪蛋白的肽 VPQLEIVPNpSAEER [M+2H]$^{2+}$ 离子的 MS-MS 谱图

在这个例子中应用的标准基本上可用来定位蛋白质中任何化学修饰。LC-MS-MS 比 MALDI-TOF 和肽质量指纹谱更适合用于蛋白质的化学修饰定位，因为 MS-MS 谱图提供肽序列信息（b 和 y 离子）和关于修饰的特定信息（如中性丢失或产物离子）。在这一章的后面将讨论这些特定修饰谱图的特征在鉴定其他蛋白质修饰中的应用。

14.4 样品富集

很明显，我们得到修饰肽的高质量 MS 或 MS-MS 谱图后，就可以鉴定肽和修饰位点，因此确定蛋白质修饰的关键是获得修饰肽的 MS 或 MS-MS 谱图。细胞中特定蛋白质的全部拷贝中只有一小部分可能带有特定修饰。例如，许多蛋白激酶底物迅速磷酸化和脱磷酸化．所以在一个特定时间一个蛋白质只存在很少磷酸化拷贝。当试图分析蛋白质样品的修饰时，样品中大多数肽是非修饰的，这在 MALDI-TOF 分析中意味着非修饰肽比修饰肽有强得多的信号。对 LC-MS-MS 分析，这意味着非修饰肽的肽离子更强，选择进行 MS-MS 裂解的可能性更大。

为解决这一问题，我们很容易想到的方法是使用富集策略增加修饰蛋白质或肽在样品中所占的比例。但在蛋白质或肽水平上，这取决于修饰的性质和丰度。相对于整个蛋白质，大多数修饰蛋白质的量很少，但修饰会改变蛋白质的某些性质。例如，改变蛋白质等电点的修饰可影响蛋白质在 2D 凝胶中的迁移，不同修饰的蛋白质可出现“点横向排列”（见第 4 章）。而肽比蛋白质小得多，肽的修饰可极大地

影响它们的行为和化学性质。对于修饰肽，富集策略必须根据修饰部分本身的化学、物理或免疫学性质。基本步骤如图 14.3 所示。含有修饰和非修饰肽的蛋白质消化物上样到含有某种固定化配基的柱上，非修饰肽对柱的亲和力较小，容易通过柱，而修饰肽与柱紧密结合。通过洗涤将非修饰肽从柱上洗脱。用破坏修饰肽与固定化配基相互作用的溶液将修饰肽洗脱。磷酸化的肽对固定化金属显示很强的亲和力，这是因为阴离子磷酸基团对多价金属阳离子的强亲和力。从蛋白质消化物中分离磷酸肽的固定化金属亲和层析（IMAC）方法就是利用了这一性质。尽管这种方法能很好地富集磷酸化肽样品，但各种磷酸肽与 IMAC 配基的亲和力会有所变化，因而使用这些柱需要很多技巧，必须仔细改进实验方案。富集样品的另一种方法是使用直接对修饰部分的固定化抗体。例如，可使用磷酸化氨基酸的抗体，通过免疫沉淀或固定化抗体柱捕获磷酸肽。和 IMAC 方法一样，这种方法取决于抗体对修饰肽的相对亲和力。对捕获生物异源物质修饰的肽（见下文），这种方法是可行的，因为这种抗体已被用于核酸加合物的类似的富集，供其后的 MS 分析。与抗磷酸氨基酸抗体一样，抗体与修饰肽的亲和力决定了富集样品的程度。综上所述，我们提出的每个富集策略都需要仔细优化和注意细节。

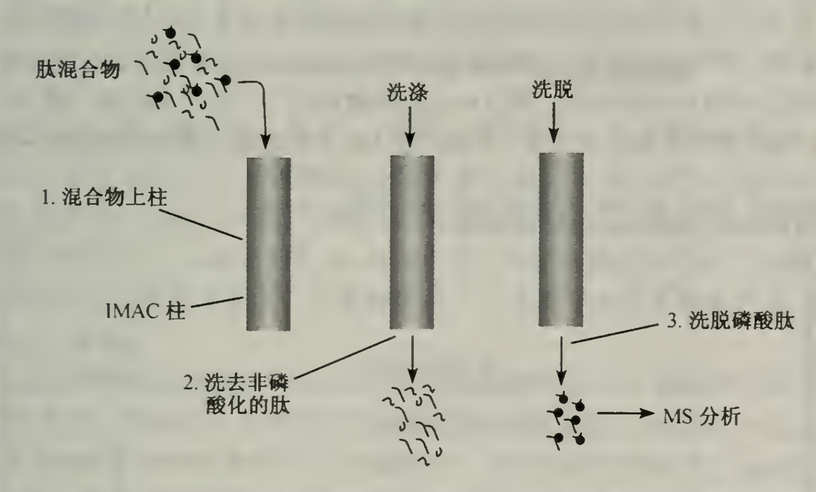


图 14.3　用固定化金属亲和层析（IMAC）从肽消化物中分离磷酸化肽

14.5　采集用于修饰研究的 MS-MS 数据

如前面提到的，一旦得到修饰肽的 MS-MS 谱图，我们就有可能推出肽序列和修饰位置。即使设计了富集修饰肽样品的策略，富集也不是完美的，仍必须在许多 MS-MS 谱图中进行挑选并确定哪一个相当于修饰肽。这与所有肽混合物的 LC-MS-MS 分析面临的困境类似：需要处理极大量的数据。幸运的是可以使用熟悉的数据缩减算法和软件工具筛选相当于修饰肽的 MS-MS 扫描数据。Sequest 程序允许使用者指定若干可能出现于蛋白质的普通低分子质量修饰。例如，使用

者可以指定，正在进行 MS-MS 谱图分析的肽中，丝氨酸、苏氨酸和酪氨酸残基的磷酸化。Sequest 将 MS-MS 数据与从数据库序列产生的虚拟 MS-MS 扫描进行关联。数据库中既包括氨基酸修饰肽也包括氨基酸非修饰肽。例如，一个 MS-MS 扫描可能对一个有丝氨酸残基的数据库序列显示很高的 Sequest 相关得分，如果对磷酸丝氨酸肽序列的相关性很高，而与非磷酸化的序列的相关性很低，这个 MS-MS 谱图可能是来自磷酸化的肽。如果对谱图中 Sequest 指定的离子检查核实，磷酸化肽的 b 和 y 离子系列有相应于磷酸化的变化，可以确定肽片段发生了磷酸化修饰。可用 Sequest 寻找各种简单、低分子质量肽修饰。这个方法很适合下列情况：①化学修饰的化学性质（质量）可以预测；②修饰导致 MS-MS 谱图产生变化；③修饰的质量是在 Sequest 和类似程序规定的范围之内。

检测蛋白质修饰的第二个方法是用第 10 章描述的 SALSA 算法分析 MS-MS 数据。许多肽修饰产生特定特征的 MS-MS 谱图。例如磷酸化的丝氨酸和苏氨酸在 MS-MS（图 14.2）中消除磷酸（98 Da）。可在谱图中观察到分别低于双电荷和单电荷前体离子 49 和 98 单位的产物离子。其他修饰也可能在 MS-MS 谱图中产生特定产物离子。例如，多环芳香烃修饰的肽随着烃的解离而裂解成为强产物离子。总之，在肽序列中任何氨基酸的稳定修饰都会改变 MS-MS 谱图中的 b 和 y 离子系列，因为修饰影响了被修饰氨基酸的表观残基质量。再来讨论在第 8 章中 AVAGCAGAR 肽的例子。图 14.4 是非修饰的 AVAGCAGAR（图 14.4A）和 S-羧甲基-AVAGCAGAR（图 14.4B）的 MS-MS 谱图。这两个肽的 y_1～y_4 离子

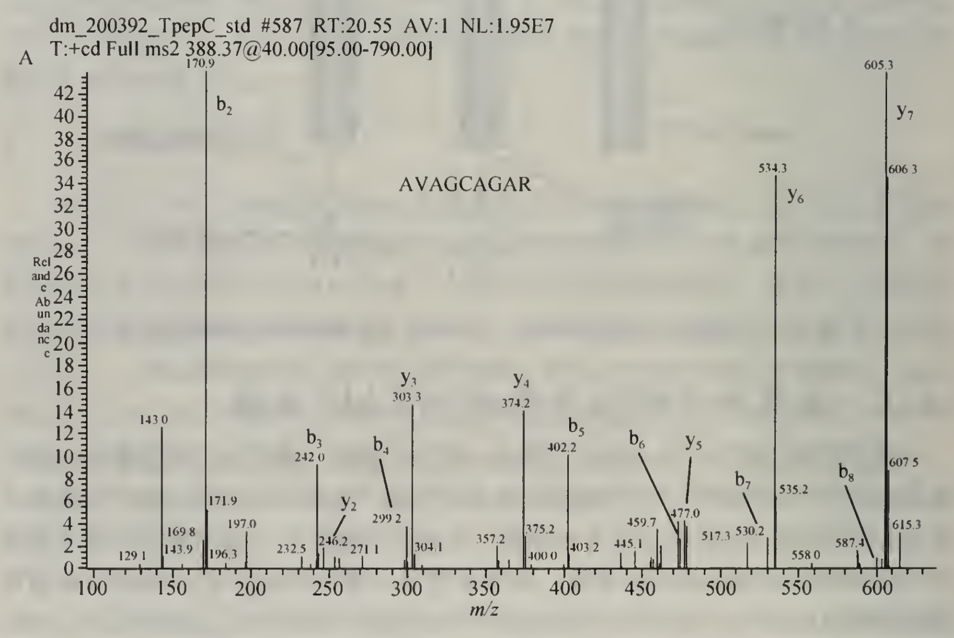


图 14.4A　非修饰肽 AVAGCAGAR 的 MS-MS 谱图

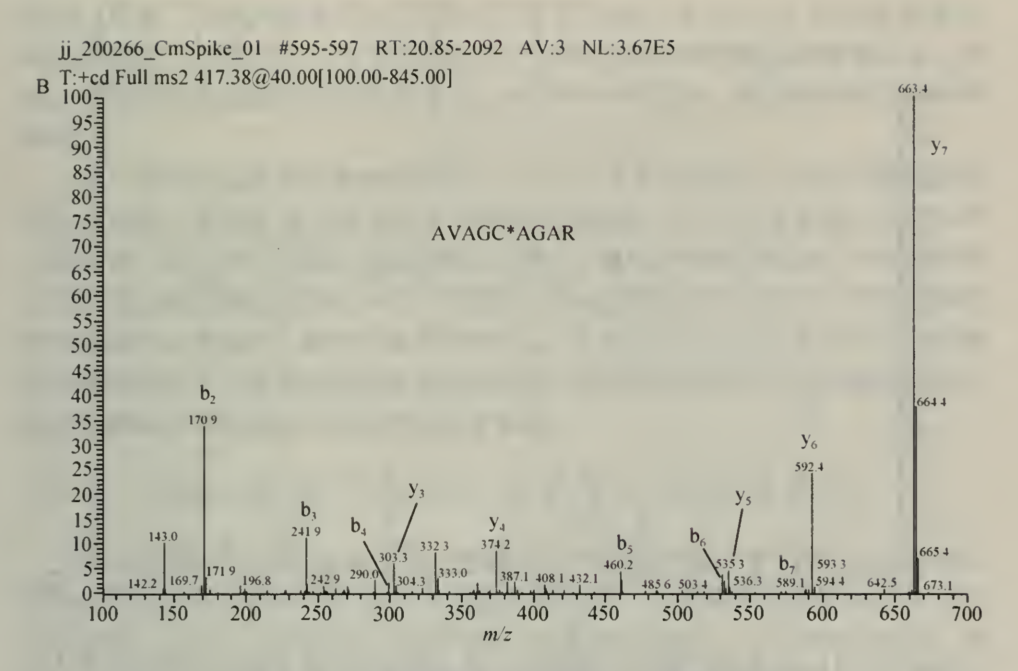


图 14.4B S-羧甲基-AVAGCAGAR 的 MS-MS 谱图

有相同的 m/z 值（未检测 y_1离子），但是 y_5离子不同。在非修饰的肽中，y_5离子 $CAGAR^+$位于 m/z 477，而修饰肽 S-羧甲基 $CAGAR^+$ 的 y_5离子位于 m/z 535，质量差异为 58，这相当于羧甲基修饰。在修饰肽中，y_5～y_8离子比非修饰肽的 MS-MS 谱图中 y_5～y_8离子都高 58 m/z 单位。相同的改变也存在于 b 离子系列。在两种肽中 b_1～b_4离子相同，但修饰肽的 b_5～b_8离子比非修饰肽的 b_5～b_8离子高 58 m/z 单位。

在第 10 章我们讨论了 SALSA 算法可检测被特定 m/z 值隔开的离子系列的 MS-MS 谱图。这种离子系列图型代表特定氨基酸基序。SALSA 产生由 b 或 y 离子系列中离子相对距离定义的“虚拟尺”，在 MS-MS 扫描中“虚拟尺”可用来检测离子系列与之匹配的谱图。对 AVAGCAGAR 肽和序列类似的肽，使用相对应于肽中间部分的“GACGA”尺进行匹配。前面的例子中，AVAG-CAGAR 肽的半胱氨酸残基上引入修饰，使分别开始于 y_5和 b_5离子的 y 和 b 离子系列发生移动。其结果是尺与修饰肽的部分 y 离子而不是与全部 y 离子系列匹配（图 14.5）。当匹配从观察到的最高 y 离子（y_7）开始时，可发现 y_7、y_6和 y_5离子与“虚拟尺”匹配，而 y_2、y_3和 y_4离子不匹配。如果从观察到的最低 y 离子（y_2，图 14.5 中未标出）开始匹配，可发现 y_2～y_4离子匹配，而 y_5～y_7离子不匹配。在这两种匹配方式中都有部分离子匹配。SALSA 对这些部分匹配的MS-MS谱图给出有意义的得分，但得分低于非修饰肽的MS-MS谱图的

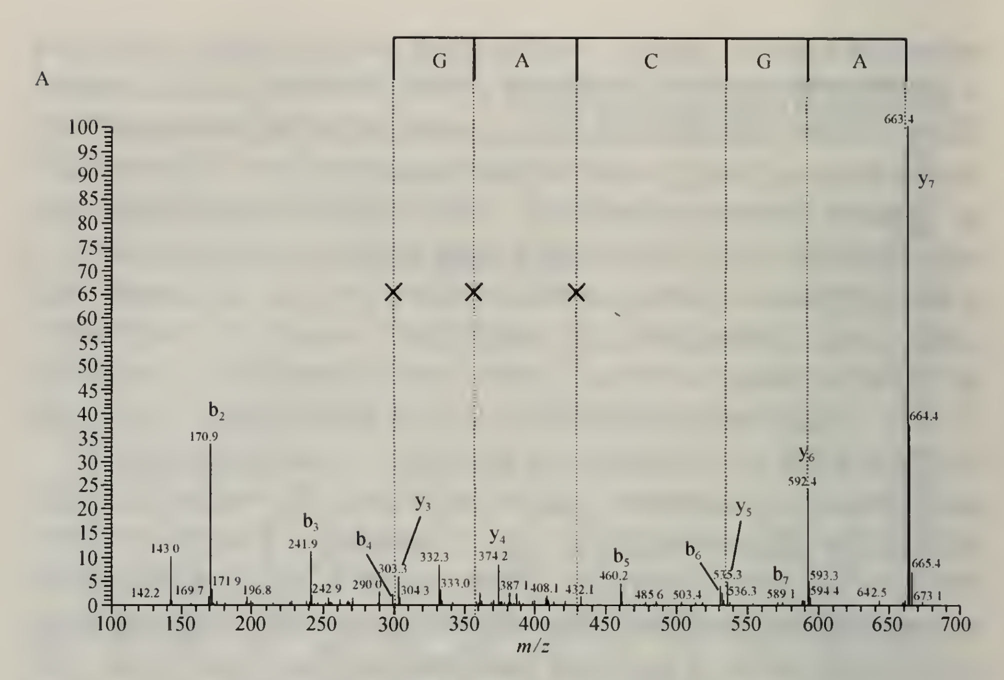


图 14.5A　用 GACGA 序列“虚拟尺”检测高 m/z 值 y 离子系列中的 S-羧甲基 AVAGC * AGAR

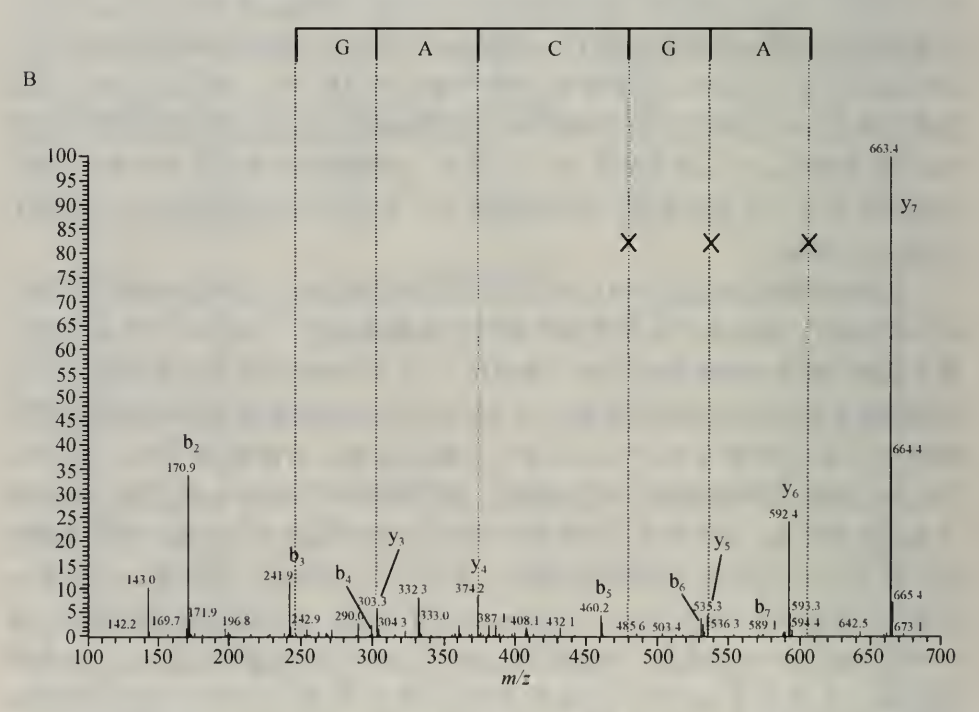


图 14.5B　用 GACGA 序列“虚拟尺”检测低 m/z 值 y 离子系列中的 S-羧甲基 AVAGC * AGAR

得分。然而，这些部分离子系列的匹配可以让我们用 SALSA 鉴定修饰肽的 MS-MS谱图。这些 MS-MS 谱图的分析可以确定肽序列和修饰的精确位置。这样，我们可以定位修饰肽在数据库蛋白质序列中的位置，确定修饰蛋白质和修饰位点。

这一基本方法是鉴定蛋白质组的有力工具。在蛋白质组中，蛋白质修饰影响蛋白质功能、蛋白质-蛋白质相互作用和蛋白质转换。SALSA 算法能区分具有序列或修饰（或二者）特征的谱图。然而 SALSA 鉴定修饰肽的前提是 MS 分析需记录下我们感兴趣肽片段的 MS-MS 谱图，因此这使我们又回到这一章前面提到的关键问题：覆盖率。最大可能地利用 Sequest 和 SALSA 方法鉴定蛋白质的修饰，需要得到尽可能多的混合物中肽的谱图，因此蛋白质消化、肽分离和仪器灵敏度达到最佳对绘制蛋白质修饰谱是重要的。

14.6 Sequest 和 SALSA 结合鉴定蛋白质修饰

Sequest 和 SALSA 单独使用时在绘制蛋白质修饰谱中都有很大的局限性。例如，我们从含有许多不同蛋白质的样品中得到了大量修饰和非修饰肽的 MS-MS 谱图。Sequest 可检测在已知氨基酸（如磷酸丝氨酸）上可预测的修饰。然而，当修饰的性质和被修饰的氨基酸不能预测时，Sequest 一般不能检测修饰形式。在这种状况下，Sequest 试图使修饰肽的谱图与数据库中未修饰的序列相匹配，从而使匹配发生错误。

同理，用 SALSA 分析这组数据会出现不同的问题。我们试图用 SALSA 寻找显示某些可预测特征（如磷酸丝氨酸和磷酸苏氨酸的磷酸中性丢失）的 MS-MS 谱图。然而，还是无法检测不能预测的修饰，或不产生明显丢失或产物离子（如磷酸酪氨酸）的修饰。如在前面和在第 10 章描述的，用 SALSA 检测修饰肽 MS-MS 谱图的最有效方式是进行序列基序检索。但要这样做必须知道检索什么基序。除非知道样品中存在什么蛋白质，否则无法这样做。

Sequest 和 SALSA 可以联用，使其更好地发挥作用（图 14.6）。基本方法如下：先用 Sequest 分析数据，许多 MS-MS 谱图数据与数据库序列成功关联。Sequest 不能正确鉴定修饰肽，但可以鉴定肽的非修饰化形式。Sequest 检索产生样品中某些标志蛋白质的目录（如图 14.6 中的蛋白质 1、2、3 和 4）。

接下来用 SALSA 检索这些蛋白质的肽所代表的序列基序。这些序列基序检索不仅仅鉴定非修饰肽的 MS-MS 谱图（这可能已由 Sequest 鉴定），而且也鉴定这样一些 MS-MS 谱图：它们与非修饰肽谱图有离子系 列同源性，也有肽质量和产物离子绝对 m/z 值的不同。这些谱图相当于修饰肽。对这些谱图进行研究可以推测每一个修饰的质量和序列位置。结合的 Sequest/SALSA 策略是使蛋白质修饰得到鉴定和定位的最完美的方式。

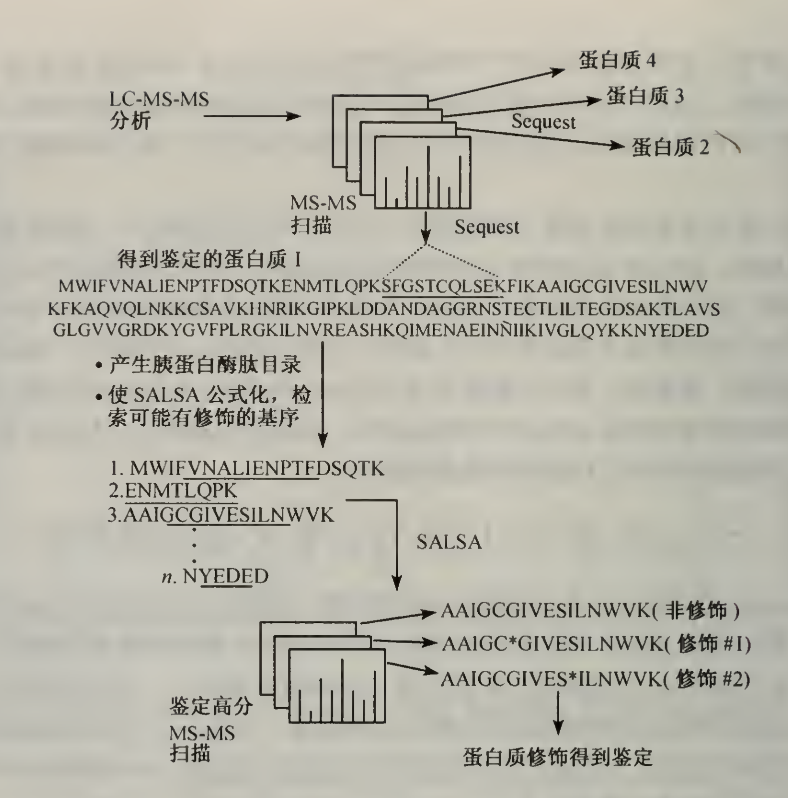


图 14.6　LC-MS-MS、Sequest 和 SALSA 联用鉴定混合物的蛋白质组分和蛋白质的修饰

推荐读物

Gatlin，C. L.，Eng，J. K.，Cross，S. T.，Detter，J. C.，and Yates，J. R.，Ⅲ（2000）Automated identification of amino acid sequence variations in proteins by HPLC/microspray tandem mass spectrometry. *Anal. Chem.* **72**，757—763.

Liebler，D. C.，Hansen，B. T.，Davey，S. W.，Tiscareno，L.，and Mason，D. E.（2001）Peptide sequence motif analysis of tandem ms data with the SALSA algorithm. *Anal. Chem.*，in press.

Neubauer，G. and Mann，M.（1999）Mapping of phosphorylation sites of gel-isolated proteins by nanoelectrospray tandem mass spectrometry：potentials and limitations. *Anal. Chem.* **71**，235—242.

Zhou，H.，Watts，J. D.，and Aebersold，R.（2001）A systematic approach to the analysis of protein phosphorylation. *Nat. Biotechnol.* **19**，375—378.

15 蛋白质组学的新方向

15.1 正在发展的技术，正在出现的工艺

当我们问在分析蛋白质组学工作中最关键的是什么时，有两件事情涌现在脑海中：灵敏度和高通量。灵敏度是重要的，因为蛋白质组学要求按照蛋白质的天然丰度分析，在许多情况下我们需要分析以极低水平存在的蛋白质。高通量也是重要的，因为要真正分析蛋白质组（与蛋白质相反），必须要尽可能快地进行大量蛋白质的分析。显然，蛋白质组学工艺正在朝高灵敏度和高通量的方向发展。

高灵敏度和高通量需要工艺和技术上的发展。“工艺”是指提供基本分析能力的仪器或仪器方法，如 MS 仪器、离子源、层析仪器等等。另一方面，“技术”是指为了从已有仪器得到最多实验结果所采取的步骤。区分这两个领域是重要的，因为这二者的提高将带动蛋白质组学未来的进步。

这本书前面描述的大多数仪器已至少存在五年了。20 世纪 90 年代早期 ESI-LC-MS-MS 就已经存在于实验室，自 1996 年以后已经得到广泛使用。在同一时间许多实验室也已使用 MALDI-TOF 仪器。在过去五年中已经广泛使用新型高分辨率的 TOF 分析器。近几年 MS 仪器性能已有了极大提高。确实，典型的 MALDI-TOF 或 LC-MS 系统比五年前出售的相同仪器的灵敏度高出一个数量级。这反映了在质量分析器工艺、ESI 和 MALDI 离子源设计、探测器灵敏度和系统电子学等方面的提高。

蛋白质组学中灵敏度的提高主要是因为用于 MALDI-TOF 和 ESI-LC-MS 的样品制备和导入技术的提高。更有效的蛋白质抽提和消化技术能从复杂样品得到更好的蛋白质和肽产量。改进的样品纯化步骤可消除去污剂和盐污染对电离和 MS 分析的干扰。新型 LC 系统较低的流量和更低的死体积保证少量样品更有效地传递到 MS 仪器。MS 仪器在低流量时往往运转最好（见下文）。从事蛋白质组学和分析蛋白质化学工作的研究人员在继续开发可提供更高灵敏度和更大通量的新技术。

上述改进将使蛋白质组分析更灵敏，并将促进蛋白质组学继续快速发展。其他正在出现的工艺会进一步提高我们分析蛋白质组的能力。在下面几节着重介绍四方面正在出现的工艺。

15.2 新型 MS 仪器

用于肽和蛋白质的 MS 分析仪器发展得很快，包括技术和工艺上的发展。尽

管质量分析器和检测器的灵敏度已有很大提高，但样品制备和导入方法的不断改进有利于检测灵敏度的进一步提高。在这些“尖端”创新中最值得注意的是用于ESI-LC-MS仪器的超低流量样品引入系统的使用。“纳喷”是描述这些技术的常用术语，样品以每分钟50～500纳升的速度流入ESI离子源（而一般通常使用的狭口柱的流速是每分钟50～500微升）。样品导入过程中降低流速可使肽离子从溶液到质量分析器更有效的转移，从而可在阿摩尔到飞摩尔范围内对样品进行MS-MS分析，而高流速分析系统所需的样品分析量要高出两个数量级。纳喷可以用融合硅毛细管进行。硅毛细管用HPLC分离介质装填，具有LC柱和电喷射针的双重功能。另一种分析方法是样品上样到空的纳喷针，在纳喷针中的肽不经串联分离直接进入MS。在串联LC-MS-MS的多维层析应用中，Yates及其同事使用的融合硅毛细管用离子交换和RP LC分离介质（见第4章）装填。在过去2～3年中，纳喷离子源已在许多蛋白质组学实验室使用。然而，现有纳喷离子源较低的精密度、不易与自动工具（如自动进样器）的联接以及仪器较低的可靠性都限制了这一方法的广泛使用。由于纳喷的高灵敏度优点，最近已引入更可靠和更容易操作的商业纳喷离子源和附件。纳喷技术的广泛使用表明它在蛋白质组学分析中将成为默认的LC-MS-MS模式。

在第6章最后提到的较新的、功能更强大的MS仪器是Q-TOF（四极杆-飞行时间）和FT（傅里叶变换）质量分析器。FT仪器可达到质量分辨率的极限，已越来越多地在分析复杂肽混合物中使用。在复杂混合物中精确和高分辨率的肽质量测定，允许使用与第7章讨论过的肽质量指纹谱不同的方式鉴定蛋白质。在FT分析中称为“精确质量标记”的已知肽质量在某些情况下可作为独特的标识。原则上精确质量标记对蛋白质组分析是非常重要的综合方法。限制FT仪器广泛使用的主要因素是昂贵的价格和相当复杂的操作。

Q-TOF仪器（大都装备有ESI离子源）在蛋白质组学工作中正在被广泛地使用。Q-TOF与离子阱和三级四极杆仪器相比其主要优点是TOF质量分析器有高质量分辨率。由于可以提供更高的分辨率和更高的质量精确性（要有适当的校准），因此Q-TOF根据肽MS-MS数据进行从头序列解释变得更容易。前体离子选择和产物离子分析的高精确性非常有利于用数据库相关算法（见第9章）从MS-MS谱图精确鉴定蛋白质序列。Q-TOF具有相当于或更高于离子阱和MALDI-TOF仪器的灵敏度。Q-TOF技术最近的扩展应用是该分析器与MALDI联用。这种结合使MALDI电离的优点与进行MS-MS分析的能力相互补充。这与不能进行真正MS-MS（见第6章）的MALDI-TOF仪器有很大不同。采用新型四极杆设计的新一代三级四极杆仪器已于最近诞生，它可在分辨率、质量精确性和灵敏度方面与Q-TOF质量分析器不相上下。

除了这些已有MS技术的提高或进一步开发，新的串联质量分析器也不断出现。其中最令人感兴趣的是TOF-TOF质量分析器，仪器使用两个不同飞行时间

分析器的串联质量分析器，供高分辨率的前体选择和产物离子的检测。与不能进行真正 MS-MS 实验的 MALDI-TOF 仪器不同，TOF-TOF 分析器在 MS-MS 分析中对前体和产物离子有极高的分辨率，具有基于 MALDI 的高通量 MS-MS 的前景。

另外，值得注意的是 MS 在分析细胞和组织蛋白质分布的“虚拟成像”方法中的应用。最近的工作表明可以将组织切片印迹转移到聚乙烯膜上，用 MALDI 基质包被，然后对印迹表面的各个位置进行一系列 MALDI 分析。一系列有序的 MALDI 激光“发射”的排列和谱图的记录，可以在整个印迹表面记录肽和蛋白质质量的谱图图型。具有特定质量的特征数据表明相关蛋白质或肽在组织切片中的空间分布。这种技术的进一步开发和最终与串联质量分析器的联用，将提供蛋白质组学与生物样品成像研究相结合的有力新工具。

15.3 自动化和机器人

在本书描述分析蛋白质组学技术时，我们集中在基础技术方面。例如，在 2D 凝胶分析中，可能选择若干蛋白质点供 MS 分析。这意味着切出每一个蛋白质点，对每个蛋白质样品进行凝胶中消化，对所产生的肽进行 MS 分析，然后用适当的软件来分析数据。尽管这种方法是可行的，但它在高通量分析方面是受限制的。而且由于在手工样品加工中不可避免的差异导致每个样品质量的不同。当然，解决这个问题的关键是在分析过程中尽可能的实现自动化操作，以增加蛋白质组分析的速度和可靠性。

在 2D 凝胶的研究中，有几家公司出售软件可用来帮助凝胶中蛋白质点的自动成像。在图型识别和比较算法的帮助下，选择用于分析的蛋白质点可在很大程度上实现自动化。然后软件驱动自动“点切割器”，收集凝胶碎片，并将碎片转移到供自动消化和为 MS 做准备的机器人装置。在许多情况下，这些机器人实际上可以将制备的样品转移到 MALDI 靶或供 LC-MS 分析的自动进样器。整个过程的自动化极大提高了蛋白质组分析的速度。通过机器人进行的可高度重复的消化和其他样品制备步骤，减少了手工样品制备时不可避免的操作差异。此外，控制这些自动化系统的软件提供自动样品追踪和有关方面的质量控制，这对高通量分析是重要的。

样品分析后的自动化有助于数据分析。例如，MS 数据文件的自动化处理能够从收集到的数据进行完全自动化的或半自动化的蛋白质鉴定。当然，评估和解释这些分析结果是我们的任务。自动化工具在样品制备、分析以及数据采集和组织等方面对蛋白质组的大规模分析是必需的。

15.4 微级和纳级仪器操作

实际上在所有技术领域中的一个重要课题是微型化。微型技术特别适用于高

灵敏度分析工作，微型技术中分析工具的大小与分析样品（细胞中的蛋白质和肽）的大小更接近。本书描述的大多数技术的低效率是因为试图在内部大小为微米或毫米的柱、凝胶和 MS 离子源分析皮摩尔或飞摩尔的肽。这如同沿街滚动几个弹子，并试图在街道的另一头全部回收这些弹子。由于肽可能与许多表面和组分相互作用，丢失是不可避免的。许多样品转移步骤（移液、层析和从凝胶中洗脱）都可能丢失肽。

进行微级或纳级分离和仪器操作的总体思路是尽量降低分析物和装置大小的差别，降低分析的无效性。最近微流装置的开发越来越受到关注，这种装置可用于抽提、消化和其他制备蛋白质和肽的步骤中。这种装置所需的样品体积在皮升到微升的范围。在公开发表的文献中已报告了大量原型装置，另有一些在进行专利商业开发。这种装置的一个共同特征是具有与用于微电路的硅芯片类似的结构，这有助于电子控制和检测器与装置的结合。

新的微型样品制备或分离装置常使用平行设计，可同时进行若干样品的处理和分析。这强调了蛋白质组分析的另一个规则：高通量。本书大部分内容中描述的蛋白质组学不能满足由微阵建立的高度平行分析的标准。多个样品的消化、分离和 MS 分析的高度平行装置可极大增加蛋白质组分析的速度。最后，微型 MS 离子源正在开发中，它可有效地使微型肽分离装置与质量分析器联用。这种离子源原型使用的电离方法包括 MALDI 和 ESI。微型离子源和纳喷一样都有高灵敏度。这些微型电离离子源可使样品离子非常有效地转移到质量分析器。

15.5 蛋白质微阵

最终与 DNA 微阵匹配的蛋白质组学当然是蛋白质微阵。然而，蛋白质微阵仍存在问题。寡核苷酸微阵技术的基础是互补序列通过 Watson-Crick 碱基配对进行杂交。而蛋白质不与互补序列杂交。寡核苷酸微阵中的靶标与探针的一一对应关系，在蛋白质组学研究中是不可行的。但仍有若干实验室正在开发通过蛋白质或肽与微阵中不同识别元素专一相互作用，进行蛋白质分析。

蛋白质有若干不同的可能识别元素，包括选择性结合较低的分子到选择性结合高度专一的分子。前者包括离子交换介质（通过在特定溶液条件下的电荷结合蛋白质或肽）和固定化金属亲和配基（它们识别某些蛋白质功能基团，如磷酸丝氨酸、磷酸苏氨酸和磷酸酪氨酸）。后者包括直接识别蛋白质的抗体。直接识别特定蛋白质序列表位的 MAb 对其靶蛋白质有高度的选择性。核酸结合子（aptamer）是蛋白质或肽的另一类高选择性识别分子。结合子是不同的寡核苷酸序列，可识别三维空间结构排列独特的氢键供体和受体，因而不同的寡核苷酸序列可能专一地与特定结构的蛋白质或肽结合。

在蛋白质组分析中许多识别元素已用在从复杂混合物中抽提特定蛋白质和肽的操作中。Ciphergen Biosystem（www. ciphergen. com）深入开发了这种方法，

该公司提供多种订制的“芯片”用于捕获蛋白质，进行 MALDI-TOF MS 分析。使用专一性较低的捕获表面可以收集多种蛋白质，专一性较高的表面化学物质（如 MAb）可以捕获单一蛋白质及其某些变化形式的蛋白质。尽管完整蛋白质的 MALDI-TOF 分析不能提供确切鉴定，但 MALDI-TOF 提供的图谱变化信息可以提供更深入研究的基础。

蛋白质识别元素微阵也可以相当于蛋白质组学中的 MS 分析，而不是仅仅为 MS 捕获蛋白质。例如，含有许多不同抗体的微阵可用来捕获各种与抗体结合的蛋白质。用这种微阵（如用荧光标记二级抗体）可提供对特定蛋白质的高灵敏和高通量的筛选。当然，这种方法的成功使用取决于抗体对靶蛋白质的专一性和亲和力、抗体吸附化学对抗体效率的影响，以及抗体与靶蛋白质结合条件的严格性。相关方法也在发展，例如随着结合子产生和鉴定技术的提高，可以预期使用高专一性结合子微阵。作用于特定蛋白质、肽或其修饰形式的结合子的大量开发可以构建用于大规模蛋白质组分析的印刷寡核苷酸微阵。

对蛋白质组学来说，微阵方法不仅仅是作为采集蛋白质组的工具。特定蛋白质微阵可用于研究蛋白质-蛋白质相互作用以及药物和其他化学或物理因素对蛋白质相互作用的影响，可以使用印在玻片或多孔板上的蛋白质微阵在特定的环境中研究蛋白质-蛋白质或蛋白质-药物相互作用。然后可用本书前面描述的 MS 工具对在单个微阵元素上的复合物成员或蛋白质修饰进行分析。

推荐读物

Baldwin, M. A., Medzihradsky, K. F., Lock, C. M., Fisher, B., Settineri, T. A., and Burlingame, A. L. (2001) Matrix-assisted laser desorption/ionization coupled with quadrupole/orthogonal acceleration time-of-flight mass spectrometry for protein discovery, identification and structural analysis. *Anal. Chem.* **73**, 1707−1720.

Chaurand, P., Stoeckli, M., and Caprioli, R. M. (1999) Direct profiling of proteins in biological tissue sections by MALDI mass spectrometry. *Anal. Chem.* **71**, 5263−5270.

Conrads, T. P., Alving, K., Veenstra, T. D., Belov, M. E., Anderson, G. A., Anderson, D. J., et al. (2001) Quantitative proteome analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ^{15}N-metabolic labeling. *Anal. Chem.* **73**, 2132−2139.

Figeys, D. and Pinto, D. (2000) Lab-on-a-chip: a revolution in biological and medical sciences. *Anal. Chem.* **72**, 330A−335A.

Lopez, M. F. (2000) Better approaches to finding the needle in a haystack: optimizing proteome analysis through automation. *Electrophoresis* **21**, 1082−1093.

MacBeth, G. and Schreiber, S. (2001) Printing proteins as microarrays for high-throughput function determination. *Science* **289**, 1760—1763.

Medzihradsky, K. F., Campbell, J. M., Baldwin, M. A., Falick, A. M., Juhasz, P., Vestal, M. L., and Burlingame, A. L. (2000) The characteristics of peptide collision-induced dissociation using a high performance MALDI-TOF/TOF tandem mass spectrometer. *Anal. Chem.* **72**, 552—558.

Zhu, H., Biolgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M. and Snyder, M. (2001) Global analysis of protein activities using proteome chips. *Science* **293**, 2101—2105.

索　引

(Q-1473.0101)

现代生物技术前沿丛书

书　名	书号 ISBN	定价
基因免疫的原理和方法	7-03-012588-6	38元
RNAi：基因沉默指南（影印版）	7-03-012685-8	65元
从基因到基因组——DNA技术概念和应用（影印版）	7-03-012477-4	46元
结构生物学与药学研究	7-03-011688-7	48元
DNA芯片和基因表达（影印版）	7-03-012248-8	29元
计算分子生物学导论（翻译版）	7-03-011493-0	36元
植物生物技术导论（影印版）	7-03-012799-4	72元
蛋白质化学与蛋白质组学	7-03-012401-4	75元
进化生物技术——酶定向分子进化	7-03-012639-4	38元
组织工程（影印版）	7-03-013407-9	55元
生物信息学	7-03-009895-1	26元
基因的自身维护与疾病的发生	7-03-013232-7	52元
生物芯片分析(影印版)	7-03-012247-X	80元
生物芯片分析(翻译版)	7-03-013661-6	88元
蛋白质组学导论——生物学的新工具（翻译版）	7-03-014258-6	30元



销售分类建议：生物医学/生物技术/分子生物学

生命科学编辑部
联系电话：010-64012501
http://www.lifescience.com.cn
e-mail:info@lifescience.com.cn

ISBN 7-03-014258-6
定 价：30.00 元